Research Statement

The explosive growth in *connectivity* and *information sharing* across a variety of electronic platforms has been accelerating the use of inferential machine learning to guide consumers through a myriad of choices and decisions. While this vision is expected to generate many disruptive business and social opportunities, it presents a number of unprecedented challenges. First, massive amounts of data need to be collected by, and transferred across, resource-constrained devices. Second, the collected data needs to be stored, processed, and analyzed at scales never previously seen. Third, serious concerns such as access control, data privacy, and security should be rigorously addressed. My research tackles these challenges (see Figure 1).

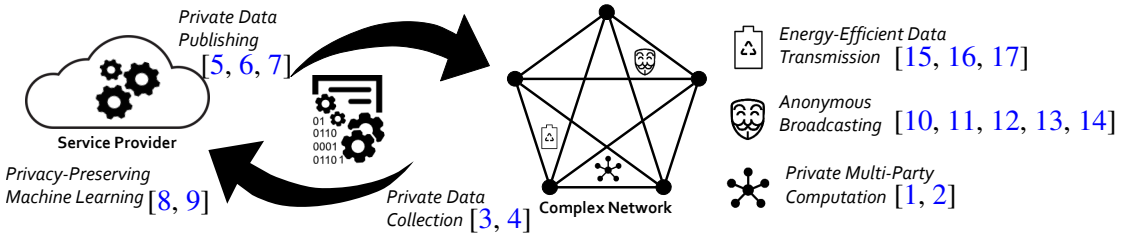

Figure 1: Designing private, scalable, and efficient information and communication networks.

Research goal. My primary research goal is to design, analyze, and implement algorithms that improve the performance, privacy, and scalability of information and communication networks.

My approach. I combine tools from information theory, statistics, machine learning, and optimization to investigate problems at the intersection of data and network sciences, computation, communication, and privacy enhancing technologies. My strongest asset is my ability to operate at both the boundaries and centers of multiple interdisciplinary fields. This is evident from the breadth and depth of my postdoctoral and doctoral research work: from privacy-preserving machine learning to anonymous broadcasting over networks to energy-efficient data transmission.

In my academic journey thus far, I have been fortunate to work with world-class researchers from Stanford, UC Berkeley, UIUC, ASU, Google Research, and Qualcomm Research. I have also had the opportunity to help my mentors write several successful NSF and DARPA research grants. My work has appeared in top-tier conferences and journals, such as NIPS, ICML, JMLR, SIGMETRICS, and Transactions on Information Theory. Further, I have received several prestigious awards, such as the Best Paper Award at ACM SIGMETRICS 2015, the Roberto Padovani Scholarship from Qualcomm Research, and the 2016 Harold L. Olesen Award for Excellence in Undergraduate Teaching from UIUC. Given my broad research background, diverse academic profile, and solid contributions, I believe that I am extremely well prepared to be an effective and engaging academic.

In the remainder of my statement, I describe my past, ongoing, and future research.

Completed Research

Privacy-preserving machine learning. In statistical analyses involving sensitive data from individuals, there is a growing tension between the need to collect and share data, and the need to preserve the privacy of individuals. The need for privacy appears in three main contexts: (1) the global privacy context, where institutions release information about individuals; (2) the local privacy context, where individuals disclose their personal information to potentially malicious service providers; (3) the multi-party privacy context, where different parties cooperate to interactively compute a function that is defined over all the parties' data. Differential privacy (DP) has recently emerged as a strong measure of privacy in all three contexts. Under DP, privacy is achieved by randomizing the data before releasing it. This leads to a natural tradeoff between privacy and utility. In a series of papers [3, 4, 5, 6, 1, 2, 8, 7], I have studied the tradeoff between DP and data utility in all three contexts, and derived optimal (and surprisingly

simple) privacy mechanisms for a variety of learning applications. Surprisingly, my findings show the universal optimality of a family of extremal privacy mechanisms called *staircase mechanisms*. This fundamental result is directly tied to the geometry of DP: staircase mechanisms are corner points of the polytope formed by the DP constraints. While the vast majority of works on DP have focused on using the Laplace mechanism, my results show that it is strictly suboptimal and can be replaced by a staircase mechanism to improve utility.

Generative adversarial privacy. My work, in collaboration with Google researchers, has shown that the strong privacy guarantees of DP often come at a significant loss in utility and sample complexity [8]. This has inspired me to search for a new privacy notion that achieves a better privacy-utility tradeoff, while still offering meaningful privacy guarantees. My search has led me to context-aware privacy. Unlike context-free notions of privacy (such as DP), context-aware approaches achieve a better privacy-utility tradeoff by exploiting dataset statistics and explicitly modeling the private variables. However, dataset statistics are hardly ever present in practice. To circumvent this issue, my ASU collaborators and I introduced a novel data-driven framework for context-aware privacy called *generative adversarial privacy* (GAP) [9]. GAP leverages recent advancements in generative adversarial networks (GANs) to allow the data holder to learn optimal privacy mechanisms directly from data. Under GAP, the privacy problem is formulated as a minimax game between two parties: (i) an adversary that wishes to learn the private variables from the sanitized data, and (ii) a privatizer (a conditional generative model) that sanitizes the dataset in a way that limits the performance of the adversary. GAP has deep game-theoretic and information-theoretic roots that allow us to provide provable privacy guarantees against a variety of strong adversaries.

Anonymous broadcasting over networks. The demand for anonymity is evident from the popularity of anonymous messaging applications. Anonymity is also crucial in nations with authoritarian governments, where the right to free expression and the personal safety of message authors hinge on anonymity. Current anonymous messaging applications hide authorship information from their users but store this sensitive information on their servers. Thus, under current designs, authorship information may be accessible to government agencies, hackers, advertisers, and of course, the company itself. In collaboration with UC Berkeley researchers, I have designed *adaptive diffusion*: a fully distributed statistical spreading protocol that uses space-and time-dependent Markov processes to broadcast messages over a social network in a way that hides authorship information [10, 11, 12, 13, 14]. Under a variety of adversarial and graph models, I have shown that adaptive diffusion enables the author to hide perfectly among the set of all users with the message. Large-scale simulations on real-life networks have verified the efficacy of this approach in practice. Anonymous communication is of broad scientific and public interest; my work on anonymous social-network messaging has received considerable of attention in my community and the media.

Energy-efficient protocols for massive wireless random access. The number of devices connected to the Internet has recently surpassed the global human population, and is projected to hit trillions within the next two decades. This super exponential growth is predominantly fueled by the emergence of low-energy, low-cost wireless devices that combine communication, computation, and sensing. These wireless devices are expected to form the fabric of smart technologies and cyberphysical systems. In this context, energy efficiency, as opposed to spectral efficiency, is extremely important because these devices are expected to be powered by tiny batteries or energy harvesting technologies. Existing technologies are designed for a very different regime: high data rates, large payloads, and abundant on-device processing power and energy. They are thus fundamentally incapable of delivering the needed energy efficiency and scalability. To address this issue, my Stanford colleagues and I have recently analyzed random access protocols that combine the device discovery and data transmission phases, embrace device collisions, trade spectral efficiency for energy efficiency, adopt simple modulation and decoding schemes. My main technical contributions are: (1) investigating the fundamental tradeoff between the maximum number of simultaneously active devices, energy efficiency, and spectral efficiency; (2) designing optimal codes that allow a massive number of energy-starved devices to sporadically access the spectrum with minimal coordination and synchronization overheads [15, 16, 17].

Ongoing and Future Research

Bringing semantics to DP. My research on differential privacy (DP) has shown that such a strong notion of privacy often comes at an unbearable cost, especially when the data lives in large dictionaries. The problem with DP is that it assumes that all the elements of the dictionary are equally sensitive. However, in practice, not all elements are equally sensitive; e.g., "google.com" vs. "marijuana.com." I am currently collaborating with Google researchers to study a generalized notion of DP that allows the privacy level to vary across the dictionary. This generalization allows system designers to put privacy where it matters without paying the cost where it does not.

Deploying GAP in the wild. My work on generative adversarial privacy (GAP) has demonstrated that under a variety of canonical dataset models, the privacy schemes that are learned from data provide rigorous privacy guarantees against a strong adversary that has access to the underlying data statistics, knows the privacy scheme, and can compute the optimal decision rule that maximizes its utility. Further, the learned privacy mechanisms match the game-theoretically optimal ones. Ultimately, my goal is to apply GAP to the 1M+ records of PG&E's massive dataset available to my Stanford mentor. In fact, PG&E is very excited about the potentials of GAP. Working towards that goal, I am currently testing GAP on the Irish Commission for Energy Regulation Smart Meter Database.

Analyzing the convergence and generalization guarantees of GAP. Adopting a data-driven notion of privacy faces important practical challenges: how to rigorously quantify the privacy guarantees of a learned privacy mechanism, and how to assess its performance relative to the theoretical optimal one? To address these fundamental questions, I plan to investigate the convergence and generalization guarantees of GAP. In the absence of such theoretical analyses, the privacy guarantees of data-driven approaches are limited to computational adversaries, making them questionable by many. While GANs have seen empirical success, their theoretical analysis has just emerged as an important research topic. I plan to use these recent advances to study the theoretical guarantees of GAP.

Designing deep learning algorithms for RF authentication systems. Many over-the-air transmissions by medical implants, wearables, and textile sensors are expected to carry sensitive data about individuals. It is therefore of utmost importance to protect such communication links against spoofing attacks. Classical cryptographic based authentication approaches are simply inadequate because they necessitate a tremendous increase in local computational, memory, and energy costs. A better approach is to harness the devices' hardware imperfections, which introduce minute (but unique) nonlinear distortions to the transmitted data signal, to authenticate devices. Such an approach shifts the complexity away from the device and places all the burden on the more capable central (receiving) node. The challenge, however, is to design robust and scalable deep learning algorithms capable of accurately classifying 10k+ devices from the data they transmit. To tackle this problem, I teamed up with a group of Stanford professors and Rockwell Collins researchers to design a deep radio authentication system. Our preliminary design is inspired by Google's FaceNet, the 2015 MegaFace Challenge winner. Moving forward, my hope is to bring machine learning to more applications in the wireless systems domain.

Bringing AI closer to the edge. The current cloud computing model requires the transfer of data from devices to the cloud where data analytics happens. However, as the number of connected devices continues to grow at an exponential rate, this model will suffer from several major limitations, such as network congestion, latency, and privacy concerns. To overcome these bottlenecks, I believe that we should push intelligence closer to the edge. While there have been several attempts to bring inference to the edge, training machine learning algorithms on the device is still largely unexplored. On-device training and inference can reduce the network traffic drastically, address latency concerns, and ensure that sensitive data never leaves the device. Nonetheless, such a novel approach introduces a few key challenges. For instance, one interesting question is: how much training should be done locally vs. globally (on the cloud)? This question requires a systematic analysis of the computation vs. communication tradeoff. Another related question is: how does one aggregate the learned models across devices? Simple averaging can be a good approach when the models are trained on the same number of i.i.d. data points. However, this is unlikely to be the case in practice. My future research will focus on addressing these fundamental questions.

References

[1] P. Kairouz, S. Oh, and P. Viswanath, "Secure multi-party differential privacy," *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[2] P. Kairouz, S. Oh, and P. Viswanath, "Differentially private multi-party computation," *Conference on Information Sciences and Systems (CISS)*, 2016.

[3] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[4] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *Journal of Machine Learning Research (JMLR)*, 2016.

[5] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *International Conference on Machine Learning (ICML)*, 2015.

[6] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Transactions on Information Theory*, 2017.

[7] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE Journal of Selected Topics in Signal Processing*, 2015.

[8] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," *International Conference on Machine Learning (ICML)*, 2016.

[9] C. Huang*, P. Kairouz*, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *To appear in Entropy*, 2017. [arXiv:1710.09549].

[10] G. Fanti*, P. Kairouz*, S. Oh, K. Ramchandran, and P. Viswanath, "Hiding the rumor source," *IEEE Transactions on Information Theory*, 2017.

[11] G. Fanti*, P. Kairouz*, S. Oh, K. Ramchandran, and P. Viswanath, "Rumor source obfuscation on irregular trees," *ACM SIGMETRICS Performance Evaluation Review*, 2016.

[12] G. Fanti*, P. Kairouz*, S. Oh, K. Ramchandran, and P. Viswanath, "Metadata-conscious anonymous messaging," *IEEE Transactions on Signal and Information Processing over Networks*, 2016.

[13] G. Fanti*, P. Kairouz*, S. Oh, K. Ramchandran, and P. Viswanath, "Metadata-conscious anonymous messaging," *International Conference on Machine Learning (ICML)*, 2016.

[14] G. Fanti*, P. Kairouz*, S. Oh, and P. Viswanath, "Spy vs. spy: Rumor source obfuscation," *ACM SIGMETRICS Performance Evaluation Review*, 2015. [**Best Paper Award**].

[15] H. A. Inan, P. Kairouz, and A. Ozgur, "Sparse combinatorial group testing for low-energy massive random access," *Submitted to IEEE Transactions on Information Theory*. [arXiv:1711.05403].

[16] I. Huseyin, P. Kairouz, and A. Ozgur, "Sparse group testing codes for low-energy massive random access," *Allerton Conference on Communications, Control, and Computing*, 2017.

[17] K. Chandrasekher, K. Lee, P. Kairouz, R. Pedarsani, and K. Ramchandran, "Asynchronous and noncoherent neighbor discovery for the IoT using sparse-graph codes," *IEEE International Conference on Communications (ICC)*, 2017.

* denotes equal contribution