Research Statement

Scott Alfeld

My research sits at the intersection of machine learning, network measurement, and security. For my PhD, I am co-advised by Professors Paul Barford and Xiaojin (Jerry) Zhu, and have been fortunate to collaborate with researchers and data scientists in and out of academia. This has given me a holistic perspective on the data analysis pipeline, leading to my publications in a broad range of venues spanning these fields and my successful work as a data scientist on real world problems. In this document I describe this perspective and outline my past and intended future contributions.

Where historically data analysis would be performed on measurements from a single, controlled experiment, many of today's datasets are aggregations of disparate but related (and potentially untrustworthy) sources. Abstractly, the pipeline of a data scientist can be modeled as follows. Some phenomenon in the real world prompts a question. We deploy sensors or use existing sensors to investigate the phenomenon. We perform analysis on the data and obtain findings and/or a predictive model. The goal is to accurately describe the phenomenon, rather than any artifacts (malicious or accidental) of the question, sensors, data, or analysis. Within this pipeline, my research focuses on inferring and incorporating knowledge of the sensors into the analysis step. Specifically, I focus on settings where a collection of **A**gents, connected via a **N**etwork, act in **T**ime (ANT settings). An "agent" is any entity which reports data, ranging from a simple sensor to a human. Within the collection of agents there may be collaborative, independent, or adversarial relationships. Note that the "network" need not be a computer or communication network, but rather can be any abstract collection of relationships between the agents.

I illustrate this flexibility with a pedagogic example of an ANT setting. Consider distributed temperature sensing at a small (e.g., within a datacenter) or large (e.g., continental) scale. The agents in this ANT setting are the thermometers, each with their own physical location, reporting temperatures over time. The network defining the relationships between their measurements is the atmosphere (sensors will show similar temperatures as those in the same jetstream, for example). In the case of measuring temperatures in a smaller area such as a datacenter, the network has an additional, virtual component based on root causes. For example if some sensors show an increase in temperature then others are likely to as well if the cause is an increase in computational load.

A broad collection of work in multi-agent systems addresses how to construct control schemes for the agents acting in ANT systems, and separately how to design ANT settings to elicit desired behavior from self-interested agents (e.g., mechanism design). Before starting at UW–Madison, my research focused on control schemes for cooperative teams of agents when the network connecting them is stochastic and unknown. In my thesis work I focus on addressing the challenges of performing machine learning and data analysis techniques in ANT settings, in particular when one or more of the agents is an adversary.

Specific Contributions

In my work I specifically address the following research questions. How should cooperative agents behave when connected via an unknown, stochastic network? How can we infer properties of the underlying network by observing the agents? How can we enhance such datasets so as to aid learners? How does one best operate, and in particular learn, when one of more of the agents is an adversary?

The Presence of an Adversary **[AAAI2017, AAAI2016, ISIT2015, ICMLWS2016]**

My work in adversarial learning is split into two tasks: detecting attacks and being robust to undetected attacks.

Toward the first goal, I collaborated with data scientists at comScore to create an efficient-at-scale method for providing theoretical bounds on the level of contamination (data corrupted or inserted by an attacker) in a dataset ([ISIT2015]). In contrast to traditional *anomaly* detection, an underlying assumption in this

contamination estimation work is that any single datapoint appears legitimate (i.e., not corrupted or inserted by an attacker) in isolation. When taken in aggregate, however, the attacker's corrupting influence on the dataset may be apparent. For example in the context of a webpage's traffic, a Firefox user running Windows may be normal, but a sudden and dramatic spike in Firefox users on Windows may be indicative of click farm traffic. Using information theoretic methods, we provide a bound on the level of contamination within a dataside given a model and desired p value. That is, our methods report the minimum number of data points which must be removed for the remaining data to match (as defined by the provided p value) the model. Importantly, our methods work at scale and are applicable to big data scenarios such as those faced by comScore, processing over 10 billion web page impressions each day.

Toward the second goal — increasing robustness of machine learning methods to undetected attacks — I focus primarily on time series forecasting. To understand how to best defend forecasters from data manipulation attacks, I first derive an adversary's most effective attack. In [AAAI2016] I derive optimal test-set attacks against linear autoregressive forecasters given a broad range of attacker constraints. My work is motivated in part by futures markets, where two parties make a bet about the future price of some commodity (e.g., corn). Both actors have an incentive to accurately forecast the commodity's price, as well as a vested interest in the other party's forecast being wrong. By subtly manipulating observations (e.g., by planting news stories about a corn shortage or hiding stores of corn), one can corrupt the other's forecast and leverage that misinformation when selecting a futures price.

While the overarching goal is to defend learners, knowledge of the optimal attacks alone is still useful. Understanding the effectiveness of attackers (or phrased differently, the vulnerability of defenders) provides a measure to be used in conjunction with model accuracy, interpretability, and other such measures when evaluating a learned model. When deriving optimal attacks, I assume a powerful attacker with full knowledge of the forecasting model, but bounded ability to manipulate input data. Given this and the attacker's target forecast (e.g., to cause a spike in price on Friday or a dip on Sunday), my methods efficiently compute the optimal perturbation within the attacker's capabilities.

In [ICMLWS2016], with continuing work in [AAAI2017], I describe a framework for defending general predictors with explicit defense actions. Available defense actions may be to limit the attacker's capabilities through legal means (e.g., by scheduling an inspection or independently verifying some values), to adjust the parameters of a learned model, or to otherwise alter the prediction process. Using this framework, I derive optimal defense strategies for autoregressive forecasters faced with *unknown* attackers. The specific goal of an attacker is assumed to be unknown to the forecaster, but the attacker uses the optimal attack for its goal. The general game-theoretic framework of the defender and unknown attackers is computationally difficult to solve. For autoregressive forecasters, and a wide class of defense actions and attackers, I reduce the problem of determining the optimal defense action to computing the spectral norm of a matrix — a computationally easy to solve task.

Controlling the Agents **[DCR2010, MSDM2011]**

In [DCR2010], I perform the first theory-driven investigation of Distributed Coordination of Exploration and Exploitation (DCEE) problems. The DCEE framework is a generalization of Distributed Constraint Optimization Problems (DCOPS). A team of agents aims to maximize their aggregate score, as in DCOPS, but the reward matrices (which define the network in this ANT setting) are stochastic and unknown. As motivation and a testbed, I used small physical robots designed to look for survivors in the remains of destroyed buildings. The team of robots aims to spread out physically while maintaining strong wireless connections with each other so as to communicate their findings with humans at a base station. How the signal strength between robots varies with their physical locations is stochastic and unknown prior to deploying the robots. Previous work had empirically observed the so-called Team Uncertainty Penalty (TUP) — where coordinating in teams leads to worse performance than agents acting independently — in common DCEE algorithms. My work ([MSDM2011]) was the first to show the TUP phenomenon on actual hardware (the aforementioned robots) and I collaborated with the mathematics department at USC to perform the first theoretical analysis of DCEE algorithms. The theoretical foundations we developed yielded two new, state of the art, distributed algorithms for DCEE problems, specifically addressing the TUP.

Inferring Underlying Properties of the Network [e-Energy2012, WWW2016]

The agents in an ANT setting may be of very different types. In the web, for example, there is a natural partition of agents: people browsing the web, and entities which provide content (including ads). When a user visits a series of web pages each placing an ad by the same third party entity, that third party can track the user via cookies (strings stored in the client's browser). In [WWW2016] I perform a measurement study of the Internet and develop a model of user information leakage — how information about a user is transmitted to third party agents via web cookies while they browse the Internet. The network in this ANT setting is defined by the users' demographic information (e.g., geographical location and interests), content demographics (e.g., product type, website genre), and the graph defined by hyperlinks between pages. With applications such as inferring user preferences in mind, I use this model to investigate how information flows across the network as a function of user type (e.g., online shopper vs social media user) and web browser cookie settings.

In [e-Energy2012], I investigate the Real Time Whole Sale Electricity Market in the Midwest USA. I perform a measurement study and derive novel methods to quantify and examine the dynamism of influence networks. In this market, the Locational Marginal Price (LMP), or price of electricity for a fixed geographic location, changes every five minutes. The different geographical locations (or "buses" in the literature) are the agents of this ANT setting, reporting local prices. The network defining the relationships between buses is a function of market forces, weather (e.g., air conditioners turning on), human activity (e.g., people moving from the city centers to the suburbs at night), and the electric power grid (defining constraints on how much electricity can flow where). As such, these relationships between locations vary over time. Two nearby cities may show very similar pricing behavior during summers or weekends, but dramatically differ from each other during winters or weekdays, for example. By quantifying these phenomena which define the similarity between agents as a function of time, my methods can be used to enhance metric-based machine learning in ANT settings.

Enhancing Data from ANT Settings [ASONAM2016, Energycon2015]

When the agents of an ANT setting are human, privacy is a concern and data is often anonymized. In [ASONAM16] I derive a method by which one may enhance an anonymized dataset while fully maintaining user anonymity rather than with de-anonymization techniques. Given an original dataset $\mathcal{U}_{\text{orig}}$ of user data, an anonymization procedure results in a dataset $\mathcal{U}_{\text{anon}}$ which is then provided to researchers. With access only to $\mathcal{U}_{\text{anon}}$, the researchers aim to infer some set of properties of $\mathcal{U}_{\text{orig}}$. My methods use knowledge of the properties one aims to infer and black-box access to the anonymization procedure to paint a more accurate picture from $\mathcal{U}_{\text{anon}}$. Specifically, I use web browsing data collected and anonymized by Cookiepedia (cookiepedia.co.uk) to preserve end-user privacy. By performing active, automated web crawling campaigns (for which user privacy is not a concern) I obtain a proxy dataset $\mathcal{S}_{\text{orig}}$. Applying the anonymization procedure to $\mathcal{S}_{\text{orig}}$ yields $\mathcal{S}_{\text{anon}}$. My methods use $\mathcal{S}_{\text{orig}}$ and $\mathcal{S}_{\text{anon}}$ to compute an approximation of the inverse transformation of the anonymization procedure (that is, the inverse of what took the unknown human data $\mathcal{U}_{\text{orig}}$ to $\mathcal{U}_{\text{anon}}$). By applying this inverse transformation to $\mathcal{U}_{\text{anon}}$, one obtains a more accurate picture of $\mathcal{U}_{\text{orig}}$ for use in inference (by human or via machine learning methods). In the cookiepedia data, for example, the anonymization procedure obscured (among other measures) the amount of traffic to each site. However, the amount of traffic obscured varied dramatically across sites, skewing the apparent popularity of different domains. In contrast with de-anonymization work, user-anonymity is maintained in full throughout the entire process.

In [Energycon2015], I create a system which allows an end user (e.g., an individual participating on the wholesale electricity market) to augment MISO's (formerly Midwest ISO, www.misoenergy.org) forecast to be more accurate for their individual needs. MISO is the Independent Systems Operator (ISO) for the electric power grid in the midwest USA; they manage the LMP markets mentioned previously, and control the flow of electricity across their footprint (spanning 15 US states and parts of Canada). MISO releases forecasts of hourly total electric load to the public for three different subregions roughly one day in advance. Individuals who produce or consume large amounts of electricity, have specific needs when it comes to forecasting regional load. For example, solar farms have a greater interest in the demand during daytime hours and entertainment media providers may care more about evening hours. I use autoregressive forecasting and

stacked generalization to reduce the error of forecasts for nighttime hours (which are intentionally obfuscated by MISO) and for short (less than 24 hours) time horizons, which is especially useful for wholesale electricity market participants. In other words, my forecasting system leverages MISO's public forecasts to provide more accurate predictions for end users. I presented my results to the Load Forecasting Team at MISO's control center in Carmel, Indiana in September 2012, where we discussed ways of using my methods to improve their own (proprietary) forecasting methods.

Short Term Future Work

Machine Teaching (MT) is the study of how to create a dataset which, when fed into a learning algorithm, results in a desired model. MT is closely related to adversarial learning — attackers aim to "teach" a learner a model which benefits them. In addition MT is applicable when the teacher's intentions are pure, for example in human education. Currently, I am collaborating with a team from the Education and Psychology departments at UW–Madison, using MT methods for optimal education design for human learners. A standard method of evaluating a model of human learning is comparing how humans and the model learn from random examples. MT provides an additional evaluation technique, where we compare how the cognitive model and humans learn from carefully constructed teaching sets. Two immediate challenges in using MT in this way are (1) in domains such as image or text classification, constructing arbitrary data points in the input space (e.g., the space of images) may lead to instances unrecognizable to humans, and (2) humans are sequential learners. Toward addressing the first issue, I am using techniques from combinatorial search and discrete optimization to construct teaching sets from a fixed candidate pool (see [SoCS2016] for preliminary work). Toward the second, my colleagues and I are extending theoretic notions of MT to sequential learners ([NIPSWS2016]).

Research Vision

The study of effective strategies in the presence of intelligent adversaries has long been of interest (cf., Tzu, circa 512 BCE) and for the foreseeable future the technological arms race between attackers and defenders will continue. I intend to continue focusing on adversarial learning which, from an applied perspective, will only increase in value as more decision making systems become automated and data driven. I will explore a variety of attack vectors which remain relatively unstudied. For example, how to attack learners which actively adapt to ongoing attacks, and attacking unknown learners and new and active areas of research. In addition, I intend to borrow from physical and computer security research to enrich our understanding of the space of adversaries. For example, I will investigate attacks based on timing (e.g., making a learner take prohibitively long to converge) and numeric stability. Analyzing such attacks will ultimately lead to the development of methods for detecting a broad class of attacks, as well as new learning algorithms which infer the underlying properties of the data despite the corrupting influence of attackers.

In addition to research directly applicable to adversarial settings, I will use MT to and adversarial learning to develop our theoretical understanding of learning. MT provides a new lens through which to view both machine and human learners, and I intend to utilize and develop this lens. Specifically, given two learners which train at the same rate to the same degree of accuracy when given *iid* data samples, MT gives us an additional way to differentiate them. And by examining the difference in optimal teaching sets for the two learners, as well as how each behaves with carefully constructed (non-*iid*) training sets, we can better understand the internal workings of the learners. With the growing popularity of deep learning methods, interpretability is becoming more challenging and more important, especially from a security point of view. From an education standpoint MT offers tools to understand how humans learn, as well as explicit methods for constructing effective teaching materials given such an understanding.

Finally, I intend to focus on two additional aspects of ANT settings: scalability and multiple views. Toward scalability, I will improve "big data" methods such as distributed learning by better leveraging knowledge of the ANT systems in which they operate. Toward multiple views, I will investigate ANT settings where each agents serve as fundamentally different sources of data (e.g., patient records and weather data during disease outbreaks). I will leverage the field of multi-view learning and the study of covariate shift to improve decision making in ANT settings.

References

[**AAAI2017**]
S. Alfeld, X. Zhu, P. Barford
"Explicit Defense Actions Against Test-Set Attacks"
in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '17)*

[**NIPSWS2016**]
X. Zhang, H. Ohannessian, A. Sen, S. Alfeld, X. Zhu
"Optimal Teaching for Online Perceptrons"
in *Constructive Machined Learning Workshop* at *NIPS*, 2016

[**ASNONAM2016**]
A. Cahn, S. Alfeld, P. Barford, S. Muthukrishnan
"What's in the Community Cookie Jar?"
in *Proceedings of the Conference on Advances in Social Network Analysis and Mining (ASONAM '16)*

[**SoCS2016**]
S. Alfeld, X. Zhu P. Barford
"Machine Teaching as Search" in *Proceedings of the Symposium on Combinatorial Search (SoCS '16)*

[**ICMLWS2016**]
S. Alfeld, P. Barford, X. Zhu
"Optimal Defense Actions Against Test Set Attacks"
in *ICML Workshop on Reliable Machine Learning in the Wild*, 2016

[**AAAI2016**]
S. Alfeld, X. Zhu, P. Barford
"Data Poisoning Attacks Against Autoregressive Models"
in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '16)*

[**WWW2016**]
A. Cahn, S. Alfeld, P. Barford, S. Muthukrishnan
"An Empiricial Study of Web Cookies"
in *Proceedings of the World Wide Web Conference (WWW '16)*

[**ISIT2015**]
M. Malloy, S. Alfeld, P. Barford
"Contamination Estimation via Convex Relaxations"
in *Proceedings of IEEE International Symposium on Information Theory (ISIT '15)*

[**Energycon2014**]
S. Alfeld, P. Barford
"Targeted Residual Analysis for Improving Electric Load Forecasting"
in *Proceedings of IEEE Energy Conference (Energycon '14)*

[**e-Energy2012**]
S. Alfeld, C. Barford, P. Barford
"Toward an Analytic Framework for the Electrical Power Grid"
in *Proceedings of the Third International Conference on Future Energy Systems (e-Energy '12)*

[**MSDM2011**]
S. Alfeld, K. Berkele, S. DeSalvo, T. Pham, D. Russo, L.J. Yan, M.E. Taylor
"Reducing the Team Uncertainty Penalty: Empirical and Theoretical Approaches"
in *Proceedings of the AAMAS workshop on Multiagent Sequential Decision Making in Uncertain Domains*, 2011

[**DCR2010**]
S. Alfeld, M.E. Taylor, P. Tandon, M. Tambe
"Towards a Theoretic Understanding of DCEE"
in *Proceedings of the AAMAS Distributed Constraint Reasoning workshop*, 2010