BIJAN MAZAHERI

github.com/honeybijan bijanmazaheri.com bmazaher@caltech.edu
Pasadena, CA (781)-985-0881

I am a computer scientist and mathematician interested in the synthesis and transportability of knowledge from and between multiple data sources. My work spans causality, mixture models, decision fusion, and distribution shift. I have experience ranging from theory to practice, such as in my website www.lacctic.com, which ranks college cross country teams using an algorithm I developed.

RESEARCH TOPICS

Decision Fusion: It is not unusual for studies and experts to disagree with each other. Such disagreement is often driven by differing contexts rather than incorrect deductions or bad data. One way to investigate contradiction is to use merged data to build a larger picture. Unfortunately, many private medical settings deny direct access to data. Through decision fusion, I seek to understand when results from different settings are in conflict and what these disagreements can indicate about the underlying system.

Confounder Identification: Phantom relationships often emerge from the combination of multiple data sources. While confounding contaminates causal relationships, it also contains the information needed to identify and remove its own impact. Assuming a bound on the cardinality of a discrete universal confounder turns the problem into a mixture model, allowing identification of within-source probability distributions. This perspective expands the notion of causal identifiability, as many graphically unidentifiable relationships can be identified.

Counterfactual Features: Features often contain a mixture of "good" and "bad" information. From a fairness standpoint, SAT scores contain information about both inherent academic ability, and also access to tutoring resources. From a domain adaptation standpoint, some information may have stable and reliable relationships with the prediction label, while other relationships break down. My work uses insights from causal inference to determine data-representations that sort between the different components of information that are hidden in these ambiguous features.

EDUCATION

California Institute of Technology - Pasadena, CA *Oct 2017 - Aug 2023*
Ph.D. Candidate
Department of Computing and Mathematical Sciences, GPA: 3.9/4.0
Awarded NSF Graduate Research Fellowship and Amazon AI4Science Research Fellowship

Cambridge University (Emmanuel College) - Cambridge, UK *Oct 2016 - Jun 2017*
Mathematics Part 1B
Supported by a Herchel Smith Fellowship
Additional classes in Computer Science and Mathematics Part II

Williams College - Williamstown, MA *Sep 2012 - Jun 2016*
Bachelor of Arts
Physics and Computer Science, GPA: 3.92/4.00
Highest Honors (Physics), Phi Beta Kappa, Sigma Xi, Magna Cum Laude
Thesis Title: RNA Macrostates and Macrokinetics

WORK EXPERIENCE

Amazon Research Causality Lab - Tübingen, Germany *Oct 2022 - Feb 2023*
Applied Scientist Intern (L5)
Worked with Dr. Michaela Hardt, Dr. Atalanti Mastakouri, and Dr. Dominik Janzing
Lead-authored a paper that was accepted to UAI 2023 and gained experience with Amazon's code review process.

BioDiscovery - El Segundo, CA *Jun 2017 - Sep 2017*
Intern
I developed methods for clustering cancers based on their genomes and implemented it within the company stack. My work has now been integrated into BioDiscovery's software and presented at a conference.

IBM T.J. Watson Research Center - Yorktown Heights, NY *Jun 2016 - Sep 2016*
Intern
Worked with Dr. Victor Kravets (mentor) and Dr. Andrew Sullivan (manager).
Projects included non-greedy and map-reduce algorithms for factoring sum of products representations. The goal of this project was to find more efficient mappings of circuits onto 2-dimensional chips.

TEACHING EXPERIENCE

Markov Chain Monte Carlo *Spring 2022*
Head TA for new class on MCMC in theoretical computer science. Developed solutions and grading rubrics for problem sets.

Physics and Mathematics *Sep 2013-Jun 2016*
TAed for undergraduate classes in Electricity and Magnetism, Classical Mechanics, Mathematical Methods for Scientists, Premed Physics, Discrete Mathematics.

PUBLICATIONS

Bijan Mazaheri, Atalanti Mastakouri, Dominik Janzing, and Michaela Hardt. Causal Information Splitting: Engineering Proxy Features for Robustness to Distribution Shifts. In *The 39th Conference on Uncertainty in Artificial Intelligence*, 2023.

Spencer Gordon, ***Bijan Mazaheri**, Yuval Rabani, and Leonard J Schulman. Causal Inference Despite Limited Global Confounding via Mixture Models. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Siddharth Jain, **Bijan Mazaheri**, Netanel Raviv, and Jehoshua Bruck. Glioblastoma signature in the DNA of blood-derived cells. *PLOS ONE* 16(9): e0256831. 2021.

Bijan Mazaheri, Siddharth Jain, and Jehoshua Bruck. Expert Graphs: Synthesizing New Expertise via Collaboration. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2447–2452, 2021.

Spencer Gordon, ***Bijan Mazaheri**, Yuval Rabani, and Leonard Schulman. Source Identification for Mixtures of Product Distributions. In *Conference on Learning Theory*, pages 2193–2216. PMLR, 2021.

Bijan Mazaheri, Siddharth Jain, and Jehoshua Bruck. Robust Correction of Sampling Bias using Cumulative Distribution Functions. *Advances in Neural Information Processing Systems*, volume 33, pages 3546–3556. Curran Associates, Inc., 2020.

* = Authorship order is alphabetical.

PREPRINTS

Bijan Mazaheri, Siddharth Jain, Matthew Cook, and Jehoshua Bruck. Combining Binary Classifiers Leads to Nontransitive Paradoxes.
Accepted IEEE ISIT 2022 but retracted due to inability to attend conference in person.

Spencer Gordon, **Bijan Mazaheri**, Leonard J Schulman, and Yuval Rabani. The sparse Hausdorff moment problem, with application to topic models. *arXiv:2007.08101*, 2020.

Siddharth Jain, **Bijan Mazaheri**, Netanel Raviv, and Jehoshua Bruck. Cancer Classification from Healthy DNA using Machine Learning. *BioRxiv*, page 517839, 2019.

Siddharth Jain, **Bijan Mazaheri**, Netanel Raviv, and Jehoshua Bruck. Short Tandem Repeats Information in TCGA is Statistically Biased by Amplification. *BioRxiv*, page 518878, 2019.

* = Authorship order is alphabetical.

PATENTS

Siddharth Jain, **Bijan Mazaheri**, Netanel Raviv, and Jehoshua Bruck. Mutation profile and related labeled genomic components, methods and systems. 2019.

PROJECTS

LACCTiC *Sep 2021 - present*
I maintain a website for collegiate cross country with 10,000 regular users based on an original algorithm for ranking performances on varying terrain. The backend runs on Python and Django and the frontend uses React, and the database is hosted on AWS.

AWARDS AND GRANTS

National Science Foundation Graduate Research Fellowship *Awarded Spring 2019*
Awarded in 2019 for a proposal to research confounding influence in causal networks.
Amazon AI4Science Research Fellowship *Awarded Spring 2022*
Funding for research with the potential to aid scientific discovery.

WORKSHOPS

Simon's Institute for Theory of Computing: Causality *Spring 2022*
4 week workshop on Causal inference methods.

TALKS

Simon's Institute for Theory of Computing *May 2023*
Title: "Causal Discovery under Limited Global Confounding"

MENTORSHIP

Caltech Cross Country Team *Sep 2018 - present*
Assistant Coach
Mentoring and supporting undergraduate students at Caltech.