Yihan Gao

Curriculum Vitae

Education

2013–present **Ph.D. Candidate in Computer Science**, *University of Illinois at Urbana-Champaign*, Urbana.
Advisor: Aditya Parameswaran

2009–2013 **B.Eng. in Computer Science (Yao Class)**, *Tsinghua University*, Beijing.

Research Interest

Database Systems (Relational DMBS in particular); Systems for Data Analysis; Automated Machine Learning System; Machine Learning Theory; Graph Embedding; Probabilistic Databases

Honors and Awards

2014 **Richard T. Cheng Endowed Fellowship**.

2009 **International Olympiad in Informatics (IOI) Gold Medalist**, *3rd place among 301 participants*.

Research Theme

My PhD research works mostly focus on using machine learning techniques to solve standard database problems: in [5], we used Bayesian Networks to compress tabular datasets; in [1], we used the Minimum Description Lengths princple to extract the structure of unidentified log datasets; in [7], we used Markov Decision Process to determine the optimal price for crowdsourcing.

Currently, I'm interested in the potential deep integrations between machine learning techniques and database systems, and my recent research works are exploring the possibility of designing a database system with integrated ML functionalities but hides away the details: in [4], we explained the correct way to interpret conditional probability estimates in agnostic settings (e.g., when the machine learning model is automatically selected); in [2], we examined the major factors that affect the performance of representation learning techniques in graph setting, which provides us with insight on how to automatically deploying such techniques in relational datasets.

Teaching Experience

Teaching Assistant for CS101 (Introduction to Programming), CS173 (Discrete Structures) in UIUC. Coach of Beijing Provincial Team for Olympiad in Informatics (2008-2011).

Publications

gaoyihan@gmail.com

[1] **Yihan Gao**, Silu Huang, and Aditya Parameswaran. "Navigating the Data Lake with Datamaran: Automatically Extracting Structure from Log Datasets". In: *ACM SIGMOD International Conference on Management of Data*. 2018.

[2] **Yihan Gao**, Chao Zhang, Jian Peng, and Aditya Parameswaran. "The Importance of Norm Regularization in Linear Graph Embedding: Theoretical Analysis and Empirical Demonstration". In: *CoRR* abs/1802.03560 (2018). arXiv: 1802.03560. URL: http://arxiv.org/abs/1802.03560.

[3] **Yihan Gao**, Aditya Parameswaran, and Jian Peng. "On the Interpretability of Conditional Probability Estimates in the Agnostic Setting". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 2017.

[4] **Yihan Gao**, Aditya Parameswaran, and Jian Peng. "On the interpretability of conditional probability estimates in the agnostic setting". In: *Electron. J. Statist.* 11.2 (2017), pp. 5198–5231. URL: https://doi.org/10.1214/17-EJS1376SI.

[5] **Yihan Gao** and Aditya Parameswaran. "Squish: Near-Optimal Compression for Archival of Relational Datasets". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1575–1584.

[6] Andris Ambainis, Mohammad Bavarian, **Yihan Gao**, Jieming Mao, Xiaoming Sun, and Song Zuo. "Tighter relations between sensitivity and other complexity measures". In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2014, pp. 101–113.

[7] **Yihan Gao** and Aditya Parameswaran. "Finish them!: Pricing algorithms for human computation". In: *Proceedings of the VLDB Endowment* 7.14 (2014), pp. 1965–1976.

[8] **Yihan Gao**, Jieming Mao, Xiaoming Sun, and Song Zuo. "On the sensitivity complexity of bipartite graph properties". In: *Theoretical Computer Science* 468 (2013), pp. 83–91.