

Guorui Xiao

Mobile: 609-373-5351

Email: grxiao@cs.ucla.edu

Website: xertxiao.github.io

RESEARCH INTEREST

I am interested in database, datastream, and machine learning system, with the ultimate goal of building scalable data-intensive systems.

EDUCATION

- **University of California, Los Angeles** Los Angeles, CA, USA
Masters of Science - Computer Science; GPA: 4.0/4.0 Expected Graduation: Mar. 2023
Advisor: Carlo Zaniolo
- **University of California, Los Angeles** Los Angeles, CA, USA
Bachelor of Science - Computer Science; GPA: 3.77/4.0; Cum Laude Graduated: Dec. 2020

PUBLICATIONS & MANUSCRIPTS

- [P1] **SoK: Revealing the Architectural Design Patterns in DDoS Defense Design Space**
Zhiyi Zhang, R. Can Aygun, **Guorui Xiao**, Sichen Song, Eric Osterweil, Angelos Stavrou, and Lixia Zhang. Under review by IEEE Symposium on Security and Privacy (**IEEE S&P**) 2023.
- [P2] **Highly Efficient String Similarity Search and Join over Compressed Indexes**
Guorui Xiao, Jin Wang, Chunbin Lin, Carlo Zaniolo. IEEE International Conference on Data Engineering (**ICDE**) 2022, pages: 232-244.
- [P3] **Demonstration of LogicLib: An Expressive Multi-Language Interface over Scalable Datalog System**
Mingda Li, Jin Wang, **Guorui Xiao**, Youfu Li, Carlo Zaniolo. ACM International Conference on Information and Knowledge Management (**CIKM**) 2022, pages: 4917-4920. (demo paper)
- [P4] **Scaling state vector sync**
Varun Patil, Sichen Song, **Guorui Xiao**, Lixia Zhang. ACM Conference on Information-Centric Networking. (**ICN**) 2022, pages: 168-170 (poster paper)
- [P5] **RaSQL: A Powerful Language and its System for Big Data Applications**
Jin Wang, **Guorui Xiao**, Jiaqi Gu, Jiacheng Wu, Carlo Zaniolo. ACM International Conference on Management of Data (**SIGMOD**) 2020, pages: 2673-2676. (demo paper)
- [M1] **A Datalog based Query Language for Supporting Recursive Query Processing over Data Streams**
[\[Link\]](#)
Guorui Xiao, Jin Wang, Jiacheng Wu, Carlo Zaniolo.
- [M2] **ReLiShare: Reliable Leaker Identification in Sensitive Dataset Sharing** [\[Link\]](#)
Zhiyi Zhang, **Guorui Xiao**, Xinyu Ma, and Lixia Zhang.

SELECTED RESEARCH PROJECTS

- **Scalable Analytics Institute (ScAi)** University of California, Los Angeles
Research Intern Dec. 2019 - Now
 - **Streaming Data Processing System that Supports Recursive Queries** [M1]
 - * Proposed a high-level query language based on Datalog for data streams to support expressing recursive queries.
 - * Devised a lightweight structure *Queue-Based Index* to avoid redundant computation and further proposed an efficient query evaluation method based on it.
 - * Designed and implemented a prototype datastream system (~15k lines of codes) to verify the effectiveness of the designs.
 - * Conducted experiments that showed we improved ~10X in throughput and ~5X in tail latency on average.
 - **Unified Compression Framework to Support String Similarity Queries** [P2]
 - * Proposed the first unified framework for offline and online construction of compressed inverted index to support String Similarity Search/Join applications to avoid expensive disk I/O costs.
 - * Devised algorithms to achieve near-optimal compression ratio in an online manner with tools like Kernel Density Estimation.
 - * Conducted experiments that showed we improved ~5X in memory consumption.
 - **Demonstration of RaSQL** [P5]
 - * Completed a demo to demonstrate that complex queries can be expressed with RaSQL and presented a user-friendly interface to interact with the RaSQL system and monitor the query results.
 - * Implemented a front end over Flask with HTML/CSS/JS, connected the front end with the RaSQL system with Py4J, prepared example queries and datasets, and contributed to the paper writing.
- **Internet Research Laboratory (IRL)** University of California, Los Angeles
Research Intern Jun. 2020 - Sep. 2020
 - **Reliable Leaker identification via shared dataset** [M2]

- * Built a prototype system focusing on Oblivious-Transfer-based end-to-end sharing that realizes reliable leaker identification and Merkle-Tree-based credential to record the resulting shared dataset.
- * Implemented a Generative-Adversarial-Network-based (GAN-based) synthetic tabular data generator to minimize the impact on the authentic shared data.
- * Prepared dataset and conducted experiments to show we achieved $< 1 \times 10^{-8}$ false negative rates by inserting only a few rows of synthetic data.
- o **Systematization of Knowledge: distributed denial-of-service (DDoS) attack [P1]**
 - * Systematically selected ~ 250 papers out of $\sim 24,000$ works related to volumetric DDoS attack and closely examined ~ 50 of them.
 - * Performed detailed analysis over selected to derive systematized repeating design patterns and a set of IP network architecture properties.
 - * Categorized the above papers into sub-categories based on their deployment locations, approaches, incentives, etc. and contributed to writing a research paper.
- o **Scaling Transport-Layer protocol in Named Data Network (NDN) [P4]**
 - * Designed and implemented both randomized and most recent partial-states States Vector Sync (p-SVS) to scale with a large number of data producers within the same group.
 - * Simulated experiments on p-SVS over an NDN simulation tool named ndnSIM over several topologies.

INDUSTRY EXPERIENCE

- **Arista Networks, Inc.** Los Angeles, CA, USA
Software Engineer Intern *Jun. 2022 - Sep. 2022*
IEEE 802.1Q Tunneling CLI
 - o Designed the new module architecture that significantly reduced the code complexity compared to the existing similar tunneling implementation and completed a detailed design document.
 - o Implemented software-side reactors and hardware-side bit setter that together can filter packets violating user-defined VLAN rules in 802.1Q tunneling. ($\sim 10k$ lines of codes)
 - o Pushed the changes to the next release to be used by all switches over a specific popular platform.
- **Taboola, Inc.** Los Angeles, CA, USA
Machine Learning/Data Science Intern *Jun. 2019 - Sep. 2019*
Knowledge Base of News Keywords
 - o Built an end-to-end pipeline with Spark SQL and Java to process data crawled by IBM Watson and further capture their embeddings with Word2Vec. ($\sim 5k$ lines of codes)
 - o Devised algorithms for de-duplicating keywords based on a combined metric, including similar neighbors, lexical similarity, etc.
 - o Proposed a Knowledge Base representation of news keywords over Neo4j to effectively visualize keywords relationships and implemented an auto-renewal process that runs daily.
- **Qihoo 360 Technology Co.** Beijing, China
Machine Learning/Data Science Intern *Jun. 2018 - Sep. 2018*
Internet Traffic Classification and Anomaly Detection
 - o Conducted surveys, implementations, and experiments on state-of-the-art machine learning algorithms for traffic anomaly detection and manually examined benign and malicious internet traffic samples.
 - o Selected features and devised an n-grams algorithm to form pseudo images from traffic.
 - o Designed a Random Forest model and a Neural Network model to achieve a 4% false positive rate and a 94% true positive rate.

TEACHING EXPERIENCE

- **COM SCI 35L: Software Construction Laboratory** Los Angeles, CA, USA
Teaching Assistant *Fall 2021*
 - o Lectured 20 hours of material focusing on Git, Shell, Vim, Java, etc., to 52 students and held 20 hours of office hours for ~ 250 students.
 - o Mentored ~ 10 groups of undergraduate students completing Node.js/React projects.
 - o Graded ~ 250 students' coding assignments and 2 exams.

MISC

- **Selected Courses:** Database System, Introduction to Machine Learning, Operating Systems, Compiler Construction, Internet Architecture and Protocols, Current Topics in Computer System Modeling Analysis.
- **Selected Languages:** Python, C/C++, Java, SQL, Bash, Datalog.
- **Selected Platforms:** Amazon EC2, Sklearn, Github, Neo4j, Apache Spark, Apache Flink, Spark Streaming, L^AT_EX.