Personal Information

- Name: Guorui Xiao Email: grxiao@cs.ucla.edu Phone: (609)-373-5351
- Address: Table 35, ScAi, UCLA, Los Angeles, CA
- Fluent in Chinese and English

Guorui Xiao is a final year student in the Computer Science Department at the *University of California, Los Angeles,* completing a Master's degree. He is actively applying for Ph.D. programs starting in Fall 2023. His research interests lie in Database, Datastream, and Machine Learning Systems, with the ultimate goal of building scalable data-intensive systems and becoming a top researcher.

Education

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

Publications

[1] ***A Datalog based Query Language for Supporting Recursive Query Processing over Data Streams***
Guorui Xiao, Jin Wang, Jiacheng Wu, Carlo Zaniolo. Under review by IEEE International Conference on Data Engineering (**ICDE**) 2023. (Research Paper)

[2] ***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. (Research Paper)

[3] ***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)

[4] ***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)

[5] ***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)

Research Experience

- **University of California, Los Angeles**
  Scalable Analytics Institute (ScAi) Dec. 2019 - Now
  I am currently working as a research intern on Database System, Datalog, and Data Stream with Professor Carlo Zaniolo. Previously I developed the first unified online compression framework for similarity join. My current work is related to designing and implementing a data stream processing system that supports recursion with high-level query language similar to Datalog. I contributed to both projects' design, implementation, and paper writings and submitted them as the first author.
- **University of California, Los Angeles**
  Internet Research Laboratory (IRL) Jun. 2020 - Sep. 2020
  I worked as a research intern on Network Security, NDN, and Network Protocol Design with Dr. Zhiyi Zhang and Professor Lixia Zhang. I implemented a prototype system that identifies leakers of the shared dataset. I also led the survey of a part of DDoS Detection papers in writing a Systemization of Knowledge paper. I contributed to the paper writings of both projects as the second author (Both are in progress). Additionally, I contributed to the research in designing a multi-path NDN forwarding strategy and improving the scalability of the NDN transport protocol.

Teaching Experience

- **University of California, Los Angeles**
  Henry Samueli School of Engineering
  COM SCI 35L: Software Construction Laboratory Fall 2021
  Teaching Assistant
  COM SCI 35L is an undergraduate course focusing on the fundamentals tools, techniques, and environments used in Computer Science. I developed content, examples, and slides for weekly discussions. I also held office hours weekly to guide students in solving problems encountered. I also mentored five groups of undergraduate students to complete the final projects and participated in grading assignments and exams.

Work Experience

- **Arista Networks, Inc.** Los Angeles, CA
  Software Engineer Intern Jun. 2022 - Sep.2022
  Arista is a network company that designs and manufactures highly-efficient network switches. I led the project of providing a more general functionality of Selective Q-in-Q to drop frames with unmatched VLAN mappings. This new feature will be installed on all the next generation of switches it releases.
- **Taboola, Inc.** Los Angeles, CA
  Machine Learning/Data Science Intern Jun. 2019 - Sep. 2019
  Taboola is an advertising company that recommends content to internet users. I led the development of the Keyword Knowledge Base project, which resulted in an efficient and accurate graph database used by the research group frequently.
- **Qihoo 360 Technology Co.** Beijing, China
  Machine Learning Research Intern Jun. 2018 - Sep. 2018
  Qihoo 360 is the largest Internet Security Company in China. I surveyed various existing State-of-art traffic classification models and further designed new Machine Learning models actively used by other group members.

Selected Projects

- **Streaming Data Processing System that Supports Recursive Queries [1]**
  - This work proposes a high-level query language based on Datalog for data streams that supports recursive queries and efficiently evaluates the input recursive programs.
  - Surveyed existing stream data processing systems like Flink and Esper and concluded that they are incapable of supporting recursive queries in a high-level language.
  - Proposed Streamlog based on Datalog to support streaming queries, gave various query examples, and proved the non-blocking semantics of the language.
  - Designed and implemented a data structure QBI that supports efficient evaluation of recursive queries by extending the Semi-Naíve Evaluation to inter snapshots.
  - Proposed a new evaluation algorithm based on the data structure and proved its complexity.
  - Designed a prototype system that parses input Streamlog queries, generates operator graphs, and evaluates the graphs against data streams.
  - Implemented the system with C++ and conducted extensive experiments.
  - Contributed to the writing of the research paper as the first author.
- **Unified Compression Framework to Support String Similarity Queries [2]**
  - This work aims to reduce the memory consumption of inverted indexes for string similarity query applications by compression techniques to fit the index size into the main memory and avoid extra I/O costs.
  - Examined through existing string similarity query frameworks and summarized a set of operations that works on the compressed indexes to support query processing directly on compressed lists without decompression.

    - Proposed the first general solution for online construction of compressed inverted index to support similarity join applications by leveraging a benefit model.
    - Implemented the framework with C++ and conducted extensive experiments.
    - Contributed to the writing of the research paper as the first author.
- **Demonstration of RaSQL [3]**
    - Built a front end with HTML/CSS/JS.
    - Connected the front end with the backend of the RaSQL system.
    - Prepared example queries and datasets.
    - Helped to finish a demo paper as the second author.
- **ReLiShare [RA at IRL]**
    - Built a prototype system based on end-to-end sharing that realizes reliable leaker identification
    - Prepared datasets, conducted experiments, and evaluated the performance of proposed algorithms.
    - Contributed to writing a research paper as the second author.
- **DDoS paper survey [RA at IRL]**
    - Examined closely over 50 papers related to DDoS traffic detection.
    - Categorized them into sub-categories based on their deployment locations, approaches, incentives and etc.
    - Contributed to writing a research paper as the second author.
- **IEEE 802.1Q Tunneling CLI [Intern project at Arista]**
    - Examined a similar tunneling implementation on an older platform.
    - Designed the new overall module architecture that significantly reduced the code complexity compared to the existing similar tunneling implementation and completed a detailed design document.
    - Implemented software-side reactors, hardware-side bit setter, and a comprehensive set of tests with C++.
    - Pushed the changes to the next release to be used by all switches of the Strata platforms.
- **Knowledge Base of news keywords [Intern project at Taboola]**
    - Built a pipeline with Spark SQL and Java to process data crawled by IBM Watson.
    - Constructed a Knowledge Base representation of news keywords on Neo4j.
    - Designed algorithms for deduplicating keywords based on a combined metric, including similar neighbors, lexical similarity, etc.
- **Internet Traffic Classification [Intern project at Qihoo 360]**
    - Conducted experiments on and examined benign and malicious internet traffic samples.
    - Selected features and designed an n-grams algorithm to form pseudo images from traffic.
    - Designed a Random Forest model and a Neural Network model to achieve a 4% False Positive rate and a 94% True Positive rate.

Courses & Skills

Courses:

- COM SCI 143 - Database System
- COM SCI 146 - Introduction to Machine Learning
- COM SCI 111 - Operating Systems Principles
- COM SCI 118 - Computer Network Fundamentals
- COM SCI 131 - Programming Languages.
- COM SCI 132 - Compiler Construction.
- COM SCI 214 - Big Data System
- COM SCI 217A - Internet Architecture and Protocols
- COM SCI 217B - Advanced Topics in Internet Research
- COM SCI 219 - Current Topics in Computer System Modeling Analysis
- COM SCI 211 - Network Protocol and Systems Software Design for Wireless and Mobile

Languages and Platforms

- Python, C++, Java, SQL, Datalog, GoLang
- Amazon EC2, Sklearn, Github, Neo4j, Apache Spark, Apache Flink, Spark Streaming