Jie Song

Contact Information	2260 Hayward St., Rm 4945 Ann Arbor, MI 48105, USA	(+1) 734-546-4113 jiesongk@umich.edu

Research Interests

Database usability, data mashup, data integration, data quality, data provenance and keyword search on structured data.

Education

University of Michigan, Ann Arbor, MI

Ph.D. Candidate, Computer Science and Engr. *Expected:* May 2019

- Thesis Topic: *Hands-off Data Integration*
- Advisor: Prof. H. V. Jagadish

M.S., Applied Statistics *GPA 3.85* May 2017

M.S.E., Computer Science Engr. *GPA 3.83* May 2016

B.S.E., Major: Computer Science Engr., Minor: Math. *GPA 3.71 Magna Cum Laude* May 2014

- member of Eta Kappa Nu (HKN)

Shanghai Jiao Tong University, Shanghai, China

B.S.E., Electrical and Computer Engineering Aug 2014

Research Experience

Research Assistant with Prof. H. V. Jagadish Nov 2014 to present

Database Group, Computer Science Engineering, University of Michigan

Integrating data sets is hard, especially for users with no technical proficiency. My goal is to enable automatic data integration. To this end, I designed and developed

- GeoFlux: an extensible data integration system that automatically joins government data based on geographic information;
- GeoAlign: a multi-reference crosswalk algorithm that approximates the distribution of aggregated data from units of one geographic type to units of another that outperforms the state-of-the-art crosswalk algorithms in both effectiveness and efficiency;

Interuniversity Consortium for Political and Social Research Oct 2016 to present

As the research community responds to increasing demands for public access to scientific data, the need for improvement in data documentation has become critical. However, the process of describing and documenting scientific data has remained a tedious, manual process and tools are lacked for documenting variable transformations in the manner of a workflow system or even a database. The Continous Capture of Metadata (C2Medatada) project is developing a Standard Data Transformation Language and a transformation engine that automate the capture and alignment of transformation operations in statistical scripts as metadata in codebook.

Research Scientist Intern with Xin Luna Dong, Xian Li and Bunyamin Sisman May 2017 to August 2017

Product Graph, Core ML, Amazon Inc.

To allow for resource-saving entity linkage for duplication-free data integration of high quality (99% precision and recall), for the construction of Lattice Product Graph, active learning is employed to reduce the labeling effort when coped with Machine Learning models. I have developed a general framework of active learning for entity matching at scale using Spark and implemented two batch-level adaptive sampling strategies leveraging the informativeness and representativeness of data for labeling. Our initial experimental results show that the framework is generic to the

model used and the sampling method chosen, and the sampling methods are more cost-saving than the state-of-the-art methods for the movie datasets.

Research Assistant with Prof. Dawn Song Nov 2013 to Feb 2014

Computer Science, University of California, Berkeley

Keystroke biometrics is believed to be one of the unique features that could help with identity authentication. It is advantageous for the task for its i) accessible and unobtrusive measurement and ii) low bandwidth transmission requirement. However, typing patterns can be erratic and inconsistent, and dependent on the type of keyboard used. We proved that the identify verification process of Coursera, which utilizes keystoke biometrics, is insecure, as an attacker can disguise as a genuine user with fake identity in three steps:

- detect the typing pattern of users of long texts and send it to the attacker;
- simulate and forge typing behaviors by learning patterns via k-nearest neighbor approach;
- and pass similarity tests for general anomaly detection of keystroke dynamics datasets provided with the forged typing behaviors .

Research Assistant with Prof. Michael P. Wellman Nov 2012 to Nov 2013

Strategic Reasoning Group, Computer Science Engineering, University of Michigan

High frequency trading (HFT) is an emerging topic in the trading market. We built an agent-based model to study how the interplay between low- and high- frequency trading affects asset price dynamics. Our main goal was to investigate whether high-frequency trading exacerbates market volatility and generates flash crashes. I contributed to

- building multiple types of trading agents to exploit their behaviors in real-world high frequency tradings;
- maintenance of an agent-based model of a limit-order book (LOB) market wherein heterogeneous high-frequency (HF) and low-frequency (LF) traders can interact;
- and exploration of the HFT features influence in generation of flash crashes and how they affect the process of price-recovery after a crash.

Engineering Experience

Team Leader/Software Developer May 2014 to Aug 2014

Product Delivery, LES Information Technology R&D Center (Shanghai), HP Inc.

Testing of duty cycle/life span of the new generation of printers with automatic document feeder (ADF) is an inevitable task for product quality evaluation. A typical enterprise-class model can go past 10k pages per month. The testing task, however, can be tedious and time-consuming if pages are fed manually under extreme lab conditions. We proposed an one-page copy mode solution to spare human efforts, and designed and built an automatic page feeder for HP AiO Laserjet life span testing.

Software Engineer Intern May 2013 to Jul 2013

Search and Discovery Team, Business Development Org., Amazon Inc.

Commodities information on the Amazon website has defects. The traditional way to detect such defects is to check against pre-defined rules. As a complement of the microscopic level approach, a macroscopic level detection strategy is expected to spot unknown and thus unusual defects without domain knowledge for warehouse data at scale (TB level). In addition to statistical modeling research, I mainly

- designed an overall distribution comparison method to model the commodity information across temporal snapshots to detect anomalies;
- implemented the method for big data with Hadoop in Amazon EMR and S3, and rank suspicious candidates by defection confidence for manual inspection;

- proved that the method is feasible to detect reasonable and unknown defects in addition to the microscopic level detector.

Software Engineer In Test Intern Dec 2011 to Feb 2012

Information Management Department, Tader Coal SCM Co.LTD

Publications

1. *GeoAlign: Interpolating Aggregates over Unaligned Partitions.* J. Song, D. Koutra, M. Mani, and H.V. Jagadish. 21st International Conference on Extending Database Technology (EDBT), In press.
2. *GeoFlux: Visualizing Data Mashup Though Automatic Geographic Join.* J. Song, D. Koutra, M. Mani, and H.V. Jagadish. Submitted, 2018. (Demo).

Teaching Experience

Student Instructor

EECS Department, University of Michigan

- EECS 280 - Programming and Introductory Data Structures Fall 2014
- EECS 281 - Data Structure and Algorithms Fall 2013, Winter 2014