Aristotelis Leventidis

leventidis.a@northeastern.edu | Personal Website | GitHub | Google Scholar | LinkedIn

EDUCATION

Northeastern University Boston, MA

Doctoral candidate in **Computer Science**, GPA 3.92/4.0 2018 – Oct. 2024 (Expected)

- My thesis focuses on the unsupervised detection and disambiguation of data values from massive table repositories
- **Relevant Coursework**: Scalable Data Management, Information Retrieval, Distributed Systems, Advanced Algorithms, Information Visualization, Human Computer Interaction

University of Michigan Ann Arbor, MI

B.S., Double Major in **Computer Science** and **Physics**, GPA 3.76/4.0 2013 – 2018

- **Relevant Coursework**: Machine Learning, Artificial Intelligence, Operating Systems, Web Databases and Information Systems, Computer Security, Computer Game Design, Computer Organization, Computational Physics

TECHNICAL SKILLS

- **Languages:** Proficient in Python, C/C++, SQL; Conversant in Java, JavaScript, HTML/CSS
- **Tools:** Linux, Docker, Git, LaTeX, Tableau, Spark, Hadoop
- **Libraries:** Pandas, scikit-learn, TensorFlow, PyTorch, NumPy, SciPy, NetworkX, spaCy, NLTK
- **Machine Learning Methods:** Linear Models, SVMs, Random Forest, Gradient Boosting, XGBoost, Generative Models, Kernel Density Estimation, DBSCAN, Topic Modelling, Time Series Forecasting, Graph Embeddings

PROFESSIONAL EXPERIENCE

Research Intern – AT&T Jun. 2022 – Aug. 2022

- Built structured story-graph views by extracting named entities and the understudied non-named entities from text.
- Used **Large-Language Model (LLM)** prompting and **NLP** techniques to extract non-named entities, integrated them using co-reference and entity resolution, and enhanced the constructed graph using sentiment extraction.
- The constructed story-graph views facilitate complex query answering not suitable for LLMs by exposing relationships not encoded in **Knowledge Bases (KBs)**

Research Intern – AT&T Jun. 2021 – Aug. 2021

- Developed a scalable, and adaptive algorithm to detect variable-sized alertable intervals in streaming data.
- Applied time series forecasting to integrate point outliers across aggregate views using kernel-density estimation (KDE), achieving top-k interval rankings with few tunable parameters.
- Improved F1-score by 50-80% and doubled the speed compared to other state-of-the-art methods.

RESEARCH PROJECTS

DomainNet – Unsupervised Detection of Homographs in Heterogenous Data Lakes (table-as-query) 2020 – Present

- Developed an unsupervised data-driven approach using **network-based centrality measures** to detect if a data value is a homograph (i.e., a value with more than one meaning).
- Clustered same meaning instances of a homograph using **DBSCAN** and **semantic similarity** measures.
- Proposed homograph detection algorithm can improve the accuracy of unsupervised domain discovery, entity matching, and semantic table search by as much as 30%.
- Awarded **Best Paper at EDBT 2021**

Thetis – Query-by-Example Semantic Table Search over Web Tables (thetis-project**)** 2020 – Present

- Enhanced **Knowledge Graphs (KGs)** by integrating them with table nodes, creating enriched **semantic data lakes**
- Designed and implemented a **scalable search framework** for semantic table search leveraging taxonomic information and learned semantic similarities via **entity embeddings** from the augmented KGs.
- The proposed algorithm retrieves more relevant tables (up to 5X higher recall) and faster (up to 17X faster response rate).
- Constructed a comprehensive evaluation corpus by mining categories and navigation links from Wikipedia Tables resulting in a dataset with 97X more queries and 956X more tables than any existing table search benchmarks.
- Paper accepted at **SIGIR 2024.**

QueryVis – Diagrammatic representation of SQL queries for better & faster understanding (queryvis.com) 2018 – 2020

- Developed and formalized a novel transformation process that converts SQL queries from first-order logic into intuitive diagrams, ensuring the diagrams are unambiguous and aid with user interpretability.
- Designed and conducted a pre-registered user-study on Amazon Mechanical Turk (AMT), demonstrating that users interpreted SQL queries 20% faster and with 21% fewer errors when using QueryVis compared to text-based SQL.
- Mentored two master students to expand the user-study.
- Received the **Most Reproducible Paper Award** at **SIGMOD 2021**.

ApproxPPR – Approximate Multi-Source Personalized Page Rank (PPR) in Knowledge Graphs 2020

- Developed an **approximation algorithm** for multi-source personalized PageRank (PPR) in knowledge graphs by aggregating single-source PPR scores.
- Empirically validated the accuracy of the approximation on both synthetic and real-world knowledge graphs.
- Demonstrated that the **approximation algorithm is parallelizable** and can improve query throughput by storing pre-computed scores for nodes using single source PPR.
- Proposed an **adaptive framework** for computing PPR in knowledge graphs, intelligently adjusting based on available system resources and historical query data to enhance responsiveness and resource management.

Shapeshifting Timelines – Evaluating the effect of timeline shapes on visualization task performance 2019 – 2020

- Designed and conducted a **pre-registered crowdsourced experiment** with 192 participants on Amazon Mechanical Turk (AMT) to evaluate user performance (accuracy and speed) across different timeline shapes and encoded data types.
- Performed statistical analysis using the Wilcoxon signed-rank test and ANOVA, revealing that linear vertical timelines enable the fastest data lookup, even for recurrent data.
- Developed timeline design recommendations tailored to user tasks and specific data encoding types, providing actionable insights for improving user interface efficiency.

PUBLICATIONS

- **Leventidis, A.**, Christensen, M., Lissandrini M., Di Rocco, L., Hose K., Miller, R. J., (2024) "A Large Scale Test Corpus for Semantic Table Search". *To appear in SIGIR 2024, pp. 13-24*
- **Leventidis, A**., Di Rocco, L., Gatterbauer, Miller, R. J., & Riedewald, M. (2023) "DomainNet: Homograph Detection and Understanding in Data Lake Disambiguation". *In ACM Transactions on Database Systems, 48(3), pp.1-40*.
- **Leventidis, A**., Di Rocco, L., Gatterbauer, Miller, R. J., & Riedewald, M. (2021) "DomainNet: Homograph Detection for Data Lake Disambiguation". *In EDBT: Extending Database Technology*.
- **Leventidis, A**., Zhang, J., Dunne, C., Gatterbauer, W., Jagadish, H. V., & Riedewald, M. (2020). QueryVis: Logic-based diagrams help users understand complicated SQL queries faster. *In Proceedings of the 2020 ACM SIGMOD (pp. 2303-2318)*.
- Di Bartolomeo, S., Pandey, A., **Leventidis, A.**, Saffo, D., Syeda, U. H., Carstensdottir, E., ... & Dunne, C. (2020). Evaluating the effect of timeline shape on visualization task performance. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-12).*
- Saffo, D., **Leventidis, A.**, Jain, T., Borkin, M. A., & Dunne, C. (2020). Data Comets: Designing a Visualization Tool for Analyzing Autonomous Aerial Vehicle Logs with Grounded Evaluation. *In Computer Graphics Forum (Vol. 39, No. 3, pp. 455-468).*
- Jin, D., **Leventidis, A.**, Shen, H., Zhang, R., Wu, J., & Koutra, D. (2017). PERSEUS-HUB: Interactive and collective exploration of large-scale graphs. *In Informatics (Vol. 4, No. 3, p. 22). Multidisciplinary Digital Publishing Institute*.

TEACHING EXPERIENCE

Northeastern University Jan. 2020 – Apr. 2020

Teaching Assistant, CS 3950: Introduction to Computer Science Research

- Facilitated engaging discussions to introduce undergraduate students to various research areas in computer science, fostering critical thinking and research skills.
- Developed and graded assignments that assessed students' comprehension of research papers and their ability to critique and expand on research problems.
- Delivered two 1-hour lectures on current research topics across different computer science subfields.