Personal background. My first taste of computer science was not the standard introduction to code via Hello World. Rather, I enrolled in an upper-level honors course called *Cyberscience* in my first term at Michigan and was in for a rude awakening when I discovered that the course's final project involved data science in Python. While I didn't emerge from the course a data scientist, my professor's enthusiasm and support encouraged me to continue in computing. Two years later I joined the Graph Exploration and Mining at Scale lab led by my adviser Professor Danai Koutra, hoping to apply the data-driven techniques I had first sampled in *Cyberscience*.

Several projects, papers, and one pending patent later, I am now a University of Michigan computer science PhD student fortunate to be supported by a first-year fellowship from Michigan and a Google Women Techmakers scholarship. To me, pursuing a PhD is the natural continuation of a highly satisfying undergraduate research experience. It is also a gateway to a professorship. After having several excellent mentors early in my career, I committed myself to giving back to the teaching community: I worked as a teaching assistant for **three Michigan computer science courses**, co-founded an **Ann Arbor-area Girls Who Code club**, and am now co-leading **Seven Mile Coding**, a programming club for Detroit youth from low-income families. I plan to continue contributing to the future of technology through research and teaching.

Relevant work. I research principled and scalable methods for temporal data mining on networks, and have applied my work to diverse domains and problems.

Research in network discovery. My **honors undergraduate thesis**, a collaboration with my adviser and neuroscientist Chandra Sripada, MD, addressed efficiently constructing a network representation (also called a similarity or correlation network) from a set of time series. Similarity graphs are studied in many fields, like transportation, economics, and neuroscience. I focused on the latter, in which neuroscientists create and analyze networks representing brain regions that "activate" in similar temporal patterns. Time series, obtained by neuroimaging procedures in the case of the brain, are traditionally converted to a network by computing correlation scores between *all pairs* of series before dropping scores below a specific threshold (Figure 1). A sparse graph in which weighted edges connect time series according to their correlation results.

Intellectual merit. Traditional network discovery is inefficient and wasteful. Many of the edges in the original network, which itself is created in quadratic time, are deleted after thresholding. This inefficiency also blocks network analysis, which is the most important part in neuroscience as it drives understanding how diseases and disorders change the brain's network organization. Motivated by the need for scalable network discovery, I **invented a new locality-sensitive hashing (LSH) family**, shown in Figure 1. LSH is a fast, probabilistic similarity search technique that requires a "true" distance measure, or *metric*, to provide theoretical

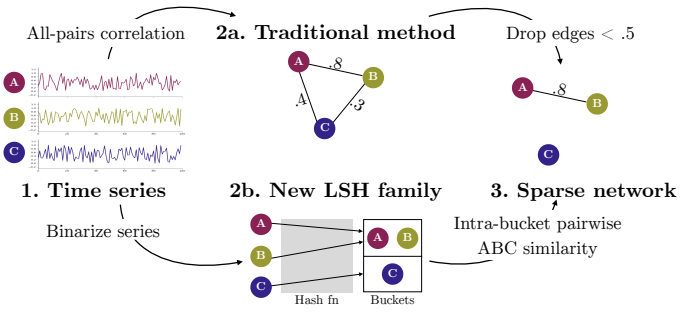

Figure 1: My proposed LSH-based method (bottom) avoids the inefficiency of traditional network discovery.

guarantees on similarity search. As correlation is not a metric, I **designed a new time series similarity measure**, Approximate Binary Correlation or ABC, and proved its corresponding distance is a metric. Unlike any other existing measure, ABC quantifies the similarity of *consecutive* fluctuations in pairs of time series.

My LSH pipeline built graphs, constructed from real brain data and larger synthetic datasets, **up to 15 times faster** than the traditional correlation-based approach. Furthermore, the predictive power of the LSH-built brain graphs matched that of the correlation-based brain graphs in a schizophrenia prediction task. I presented and defended these findings in my thesis, which received High Honors, and submitted a full conference paper accepted to the **2017 IEEE International Conference on Data Mining (ICDM)**, the premier data mining conference. One ICDM reviewer commented that my new LSH family and metric are **likely to be used by other researchers**.

Broader impacts. I presented my work several times to **demonstrate its applicability** and **increase the visibility of undergraduate research**. As a guest lecturer in Michigan's undergraduate Discrete Mathematics course, I discussed how improvements to network discovery speed can lead to **faster identification of mental disease** and provided guidance about involvement in undergraduate research. At Google (below), I outlined the applications of my LSH family, leading the way to its **use in Google's production systems**.

Research in network anomaly correlation. As an intern at Google, I researched **new machine learning approaches for understanding Google's massive network infrastructure** in collaboration with my mentor Xiang Wang. Our goal was to detect correlated network anomalies, like latency spikes and physical link failures, in linear time. Google's network generates enough data—millions of edges over many years—to make human correlation of network faults, otherwise known as "dashboard gazing", extremely difficult.

Intellectual merit. My solution, **for which a patent is pending**, consisted of two steps. The first is a variant of the LSH pipeline I proposed in my thesis. The second step is a **novel linear-time generative model**, extending an algorithm that has been used almost exclusively in the natural language processing domain, for efficiently discovering a provable underlying joint probability distribution of network anomalies.

Broader impacts. My solution has been integrated into production systems that **guide Google's network planning and operation workflows**. Other network teams at Google use our methods to simulate network planning and assess risk thereof. Working at Google showed me an extreme instance of the potential scale of impact of my research: Google's production network handles billions of queries per day. My work improved both automated and manual network monitoring, directly affecting worldwide information access.

Research in network summarization. I worked with my adviser and her PhD student to summarize a large graph, or create a smaller, interpretable model of the graph, using the Minimum Description Length (MDL) model selection criterion. Improving upon my adviser's previously introduced "Vocabulary of Graphs" (VoG) graph summarization pipeline, we scaled a quadratic-time algorithm to compose a concise graph summary.

Intellectual merit. I **designed an iterative parallel model selection heuristic** for VoG. In this algorithm, a coordinator process iteratively assigns partitions of the input graph to several threads and evaluates the threads' results until convergence is reached. During an iteration, each thread identifies a candidate subgraph to improve the global summary

model using the VoG MDL objective function and its partial knowledge of the graph. My algorithm maintained the quality of output while **speeding up our experiment runtime from over one month to several days**. My work led to us submitting a paper accepted at the **KDD 2016 Workshop on Mining and Learning with Graphs**, as well as me receiving the University of Michigan computer science department's **Outstanding Research Award**. Furthermore, to advance knowledge on the importance and methodologies of graph summarization, I **co-led a graph summarization survey** undergoing a second round of journal reviews and **led a book chapter on related techniques** to be published in a 2018 CRC Press book on advanced social media analytics.

Broader impacts. Since our work was published in a KDD workshop, I attended the conference and its associated **Broadening Participation in Data Mining (BPDM) workshop**, which provides mentorship and connections to students of backgrounds underrepresented in the data mining community. At BPDM, I shared my perspectives with other participants on the differences between academia and industry, having worked in both, and outlined my graph summarization research to give and receive advice about future directions.

Broader impacts and future goals. Beyond research, I use education to positively impact those around me. I am a 2017 recipient of the University of Michigan **Marian Sarah Parker Prize** and the **Google Women Techmakers** scholarship (formerly known as the Google Anita Borg scholarship) for my contributions to the community of women in technology. I co-founded the University of Michigan Women in Science and Engineering's **(WISE) Girls Who Code club** to help increase female representation in computing, which is currently very low. We succeeded in that several club graduates are studying computer science in college, and one is now a Girls Who Code instructor. Recently, I ran into a former student for whom I was a TA, who told me that she continued in computer science because of me. As **Seven Mile Coding**, my current teaching initiative, begins in a low-income Detroit neighborhood, I will continue to improve the culture of computing for the future. For example, I recently submitted a budget request to the Google Women Techmakers scholarship program to help fund Seven Mile Coding, which was approved.

As a researcher, I will continue to apply my work to the high-impact area of neuroscience. According to the NIH, mental disorders were the largest driver of healthcare expenditure in 2016 with a cost of $2.5 trillion, and that cost is expected to rise to $6 trillion by 2030. As I discuss in my Graduate Research Plan, I want to contribute to understanding mental disorders—especially depression, which has directly affected many of those close to me—from a data-driven, network-theoretical perspective. Furthermore, as demonstrated by my time at Google, my research is broadly applicable to domains and problems as diverse as large-scale network monitoring and anomaly detection. I hope to lead partnerships between academia and industry to solve such global-scale problems in the future.

The NSF fellowship will contribute to my career and educational goals by helping me grow my academic and professional network, attend conferences and workshops, and pursue the research opportunities that interest me most. In particular, attending conferences accelerates the transfer of knowledge. I attended KDD and BPDM with the help of a travel grant, where I learned about the state-of-the-art in my field and made many connections and friends. The NSF fellowship would help me rapidly pursue my research, teaching, and outreach goals.