In this talk, we review two of our most recent designs, Graph of Thoughts and ProbGraph, that harness graph computing for more powerful and parallelizable Artificial Intelligence and Data Analytics.

Graph of Thoughts (GoT) is a framework that advances prompting capabilities in large language models (LLMs) beyond those offered by paradigms such as Chain-of-Thought or Tree of Thoughts (ToT). The key idea and primary advantage of GoT is the ability to model the information generated by an LLM as an arbitrary graph, where units of information ("LLM thoughts") are vertices, and edges correspond to dependencies between these vertices. This approach enables combining arbitrary LLM thoughts into synergistic outcomes, distilling the essence of whole networks of thoughts, or enhancing thoughts using feedback loops. We illustrate that GoT offers advantages over state of the art on different tasks, for example increasing the quality of sorting by 62% over ToT, while simultaneously reducing costs by >31%. We ensure that GoT is extensible with new thought transformations and thus can be used to spearhead new prompting schemes. This work brings the LLM reasoning closer to human thinking or brain mechanisms such as recurrence, both of which form complex networks.

ProbGraph is a graph representation that enables simple and fast approximate parallel graph mining with strong theoretical guarantees on work, depth, and result accuracy. The key idea is to represent sets of vertices using probabilistic set representations such as Bloom filters. These representations are much faster to process than the original vertex sets thanks to vectorizability and small size. We use these representations as building blocks in important parallel graph mining algorithms such as Clique Counting or Clustering. When enhanced with ProbGraph, these algorithms significantly outperform tuned parallel exact baselines (up to nearly 50x on 32 cores) while ensuring accuracy of more than 90% for many input graph datasets. Our novel bounds and algorithms based on probabilistic set representations with desirable statistical properties are of separate interest for the data analytics community.

bio: Maciej Besta leads research on sparse graph computations and large language models at the Scalable Parallel Computing Lab (SPCL) at ETH Zurich; he also works on network topologies and occasionally other aspects of the high-performance computing landscape. He received his PhD from ETH Zurich in 2021. He published around 50 peer-reviewed scientific conference and journal articles at top venues, receiving Best Paper Awards & Nominations at ACM/IEEE Supercomputing 2013, 2014, 2019 (for two papers) and 2022 (for two papers), at ACM FPGA 2019, at ACM HPDC 2014, 2015 and 2016, and ACM Research Highlights 2018. He won, among others, the competition for the Best Student of Poland (2012), the first Google Fellowship in Parallel Computing (2013), the ACM/IEEE-CS High-Performance Computing Fellowship (2015), the ETH Medal for the Outstanding Doctoral Thesis (2021), the IEEE TCSC Outstanding PhD Dissertation Award (2021), the SPEC Kaivalya Dixit Award for Distinguished Dissertation in Performance Evaluation (2022), and the ACM SIGHPC Dissertation Award (2022). He is also a Fellow of The Explorers Club (2022). More details: https://people.inf.ethz.ch/bestam/