hausdorff CENTER FOR MATHEMATICS

UNIVERSITÄT BONN

Workshop

"Massive data Models and Computational Geometry"

September 23 – 27, 2024

organized by

Mark de Berg, Anne Driemel, Robert Krauthgamer, Morteza Monemizadeh

Abstracts

Jeff Phillips (University of Utah)

The Versatility of Mergability for Geometric Summaries on Massive Data

Abstract: The concept of mergable summaries was introduced in PoDS 2012 as a framework for forming small summaries (coresets & sketches) of massive data with guarantees on their accuracy eps as a function of size s. The key idea considers data X split into two sets X_1 and X_2, where we create summaries S_1 and S_2, each with size s and accuracy eps. Mergability refers to the ability to create a summary S of all of X by only accessing S_1 and S_2, and have the size and accuracy of S matching the s and ϵ, wrt X, of the summaries of the subsets. In this talk will have two main parts:

1. Review some geometric problems where input X are points in R^d for which we can create mergable summaries.
2. Overview the diverse big data settings where having mergability implies efficient algorithms.

Along the way, I will aim to scatter various open problems in both geometry and massive data modeling.

Pankaj K. Agarwal (Duke University)

Algorithms for Optimal Transport in Discrete and Semi-Discrete Settings

Abstract: Given a d-dimensional continuous (resp. discrete) probability distribution A and a discrete distribution B, the semi-discrete (resp. discrete) optimal transport (OT) problem asks for computing a minimum-cost plan to transport mass from A to B; we assume n to be the number of points in the support of the discrete distributions. The cost of the OT plan is referred to as the

Wasserstein or earth-mover's distance. This talk presents efficient algorithms for computing both discrete and semi-discrete OT and its variants. It also discusses algorithms for clustering under Wasserstein distance.

Jie Gao (Rutgers University)

Differentially Privacy and Discrepancy on Shortest Paths

Abstract: We consider differentially private range queries on a graph where query ranges are defined as the set of edges on a shortest path of the graph. Edges in the graph carry sensitive attributes and the goal is to report the sum of these attributes on a shortest path for counting query, using differential privacy to ensure protection of the sensitive edge attributes. Our goal is to develop mechanisms that minimize the additive error of the reported answers with the given privacy budget. I will report tight error bounds of roughly $O(n^{1/4})$ for this problem. The lower bound comes from hereditary discrepancy of shortest paths. Namely, the (vertex/edge) discrepancy of shortest paths considers coloring the vertices/edges by $+1$ or -1 and asks for minimizing the maximum magnitude of summation along any shortest path. We show nearly tight upper/lower bounds on hereditary discrepancy of $n^{1/4}$ for shortest paths, improving a previous lower bound of $\Omega(n^{1/6})$ on a classical point-line system of Erdos. This is joint work with Greg Bodwin, Chengyuan Deng, Gary Hoppenworth, Jalaj Upadhyay and Chen Wang, published in WADS'23 and ICALP'24.

Erik Waingarten (UPenn)

Average-Case Sketches

Abstract: Sketching vectors refers to the following task: compress two vectors so as to approximate their pairwise distance. Over the past decades, we've developed an almost complete understanding of sketching vectors in various cases — for the ℓ_p norms, we have tight space-approximation tradeoffs for what is achievable. Can we design better sketches if vectors come from a distribution? We introduce average-distortion sketches to address the above question, and show that better space-approximations are achievable for the ℓ_p norms. We'll discuss connections to data-dependent locality-sensitive hashing as well as average-distortion embeddings, and highlight algorithmic applications.

Sepideh Mahabadi (MIT)

Approximate Nearest Neighbor Search and Its Many Variants

Abstract: In this work I will survey some of the variants of Approximate Nearest Neighbor (ANN) search that are motivated by recent applications. For example we show how to report search results that satisfy diversity, fairness, or differential privacy.
I will then focus on incorporating diversity into ANN, and show how to use the notion of robust composable coresets to get almost matching space and query time as the standard ANN. Composable coresets are small subsets of data sets such that their union contains an approximately good solution for the whole data, with respect to an optimization task (here "diversity maximization"). These notions have further applications beyond nearest neighbor to streaming and distributed settings.
Throughout the talk, I will mention several open problems both in the context of Nearest Neighbor Search and Composable Coresets.

David Woodruff (CMU)

A Strong Separation for Adversarially Robust L_0 Estimation for Linear Sketches

Abstract: We give the first known adaptive attack against linear sketches for the well-studied L_0-estimation problem over turnstile, integer streams. For any linear streaming algorithm A that uses a sketching matrix of dimension r by n, this attack makes $O(r^8)$ queries and succeeds with high constant probability in breaking the sketch.
Joint work with Elena Gribelyuk, Honghao Lin, Huacheng Yu, and Samson Zhou.
