Research Statement

Chengyuan Deng

My research is dedicated to building the theoretical foundations for the new arenas of modern algorithm design. The rapid emergence of massive, complex and sensitive data has created a new frontier of computational challenges, demanding solutions beyond the traditional models and analytical methodologies. From the data-informed algorithmic perspective, the challenges naturally unfold through the following sequence of stages:

- **Access to massive data.** Modern datasets are produced at unprecedented scale and speed, often in streaming, dynamic, or distributed settings. The central task is to develop algorithms that can interact with such data efficiently, making useful computations possible within tight resource constraints.
- **Understanding the inherent structure.** Large datasets are complex but often shaped by underlying algebraic, combinatorial, geometric, or probabilistic patterns. Revealing and formalizing these structures is key to designing algorithms that are effective, interpretable, and widely applicable.
- **Responsible release of information.** Algorithms transform data into knowledge, which in turn raises the need to carefully regulate what is disclosed, particularly in the presence of sensitive content. This calls for frameworks that safeguard trustworthy considerations, while releasing reliable information with rigorous guarantees.

During doctoral study, my work confronts challenges along this pipeline, grounding *scalability, explainability, data geometry, and privacy* within the framework of algorithm design. First, towards **learning under limited access**, my research explores streaming and property testing models on the problem of **correlation clustering**. Our algorithms achieve a favorable approximation with resources significantly smaller than the input size. Second, my research aims to understand the role of **data geometry** in fundamental problems such as **dimension reduction, nearest neighbor search** and design algorithms tailored to the geometry. Further, my research borrows tools from combinatorial discrepancy theory to understand the **structural complexity of shortest paths**, which turns out to have an impact on **differential privacy**, falling into the regime of responsible release. Finally, under the umbrella of differential privacy, my research investigates a few problems such as **range query** and **hierarchical clustering**.

Efficient Algorithms for Structural Balance and Correlation Clustering

In correlation clustering, we are given a complete graph with binary labels. The objective of clustering is to partition the vertices into clusters (size unknown) such that the number of negative edges across clusters and positive edges inside the same clusters are minimized. If the number of clusters is fixed as $k = 2$, this is known as structural balance from social science community.

Streaming Algorithms for Structural Balance. We develop streaming algorithms on two tasks: (i) detecting whether a given graph is balanced, and (ii) finding a partition that approximates the clustering objective. The goal is to use limited resources on storage. For the first, we employ pseudorandom generators for low-degree polynomials, which require only $O(\log n)$ memory; For the second, we achieve $(1+\epsilon)$-approximation with $\tilde{O}(n)$ memory by improving the Giotis-Guruswami algorithm [GG05] and simulation with graph streams.

Property Testing for Correlation Clustering. In property testing, we consider the query model with access to the adjacency matrix. The goal is to determine if the input graph admits a 0-cost clustering solution or ε-far from it. Our main result is a simple $\Theta(1/\varepsilon)$ query complexity algorithm for structural balance, and $O(1/\varepsilon^2)$ query complexity for correlation clustering.

The related publications are recognized by **Random** [AAD+23] and in submission.

Dimension Reduction Beyond Euclidean Geometry

Traditional dimension reduction techniques have two assumptions: (i) the data points lie in Euclidean space, and (ii) the coordinates of all points are available. Many real-world applications do not satisfy these two conditions. First, the geometry is often non-Euclidean such as cosine similarity, Jaccard index, etc. Second, high-dimensional coordinates may be unavailable or costly to obtain, whereas pairwise distances are easier to access. Therefore, we study dimension reduction techniques to address these challenges. We consider the input is a symmetric hollow dissimilarity matrix D, with only two assumptions for the dissimilarity measure: $D_{ij} = D_{ji}$ and $D_{ii} = 0$.

Non-Euclidean Johnson-Lindenstrauss Lemma. JL lemma states that random linear projection can achieve dimension reduction in Euclidean space, while preserving pairwise distances up to $(1 \pm \varepsilon)$ factor. We study JL transform in the above setting and present two approaches. In both approaches, we generalize the Euclidean norm ℓ_2 to to capture the data geometry, together with parameters indicating how far it deviates from Euclidean geometry.

The first approach captures D as vectors in **pseudo-Euclidean space**. Here the distance between two vectors x, y is given by a bilinear form of signature (p, q): $\langle x, y\rangle_{p,q} = \sum_{i=1}^{p} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i$. When $p = n, q = 0$, it is the squared Euclidean norm. We interpret the parameter p resembling and q negating the Euclidean space. Our result is *fine-grained*, indicating D_{ij} can be preserved with at most $1 \pm \varepsilon \cdot C_{ij}$ multiplicative factor, where C_{ij} is the ratio of the squared Euclidean distance and squared (p, q) distance.

Our second approach is via **power distance**. We prove D can be written as a generalized power distance matrix of n points $\{p_i\}$ with the same radius $r = \sqrt{|e_n|}/2$, where e_n is the smallest eigenvalue of the Gram matrix of D. We can then apply the JL transform on the ball centers (i.e. $\{p_i\}$). The new JL lemma has an extra additive term of $4\varepsilon r^2$, with $(1 \pm \varepsilon)$ multiplicative factor preserved for every D_{ij}.

Non-Euclidean Multi-dimensional Scaling. MDS is another widely-used dimension reduction method. The algorithm selects top k largest eigenvalues of the Gram matrix to produce low-dimensional vectors. We extend MDS to the same setting above, and prove an optimal algorithm of selecting both negative and positive eigenvalues, with the aim of minimizing the "distortion". The approach is also based on *pseudo-Euclidean* bilinear forms.

Locality Sensitive Hashing in Hyperbolic Space. LSH is directly motivated from the approximate nearest neighbor search (ANNS) problem [IM98] and has broad applications. It ensures that close pairs of points are likely to be hashed into the same bucket, while distant pairs are not. It can be interpreted as a dimension reduction technique to discrete 1D, but with a relaxed goal. LSH has been well-studied in Hamming space and Euclidean space [DIIM04, AI08, MNP06, OWZ14]. We extend it into Hyperbolic space. The performance of LSH is governed by ρ, which gives query time $O(n^\rho)$ for c-approximate NNS. We design a Hyperbolic LSH with upper bound $\rho \leq 1/c$ and a lower bound of $\rho \geq 1/c^2$. The Euclidean LSH has $\rho \approx \Theta(1/c^2)$.

The related publications are recognized by **Neurips** [DGL+24, DGL+25] and in preparation.

The Discrepancy of Shortest Paths

Combinatorial discrepancy theory aims to express the complexity of a set system. We study the discrepancy of shortest path systems to understand its structural property. We show that any system of unique shortest paths in an undirected weighted graph has hereditary discrepancy $\tilde{\Theta}(n^{1/4})$, on the lower bound side improving upon $\Omega(n^{1/6})$ [CL00]. Meanwhile, we demonstrate an arbitrary path system has hereditary discrepancy of $\tilde{\Theta}(\sqrt{n})$, therefore showing a clear separation between arbitrary paths and shortest paths.

The discrepancy result demonstrates the inherent structural complexity of shortest paths, and interestingly sheds light on other applications. An immediate implication is on two problems in differential privacy: DP All Pairs Shortest Distances and DP All Sets Range Query. The discrepancy lower bound implies an $\tilde{\Omega}(n^{1/4})$ lower bound of additive error on two problems, which is currently the best known lower bound.

The related publications are recognized by **WADS** [DGUW23], **ICALP** [BDG+24] and **ITCS** [ABD+24].

Empirical Works on Machine Learning, Datasets, LLMs

I collaborate closely with industry and care about the empirical performance of algorithms.

OpenFWI: First Open-source platform for Full Waveform Inversion. Full Waveform Inversion is a technique in Geoscience to understand the subsurface structure. In recent years data-driven methods are developed for the task but almost no public datasets. We build a large-scale public platform [DFW+22, FWD+23] that provides datasets, tutorials, baseline implementations and a Kaggle competition to facilitate future research. They have significant impact in the AI for science community.

Change Point Detection. Online CPD aims to identify abrupt changes in multivariate time series. Accuracy and efficiency are the key desired properties. We develop a simple algorithm [DCZ+24] inspired by the Riemannian geometry of correlation matrices. It is able to detect changes on both marginal and joint distribution.

Large Language Models. With collaborators, I conduct surveys of LLMs on domain specialization [LZL+23] and ethical issues [DDJ+25]. The former received attention from the U.S. presidential annual report in 2024. Concrete technical projects are ongoing, including privacy audit of LLMs and cognitive obstacles in VLMs.

The related publications are recognized by **Neurips** [DFW+22, FWD+23], **CIKM** [DCZ+24], Journal of computing surveys, Journal of AI and Ethics, and in submission.

Future Research Directions

I envision a future of algorithm design shaped by principles that unify classical theory with the challenges of modern computation and data. I see them as mutually reinforcing. Therefore my future research is on this playground focusing on two topics: **Theoretical foundations of learning problems** and **Graph algorithms**. Both lines of research consider the challenges about data in its three stages as discussed at the beginning of this statement.

- **Theoretical Foundations of Learning Problems.** I view this direction with two major aspects: (1) learning problems without an established theoretical framework (2) learning problems with theoretical frameworks (e.g., k-clustering, PCA, SVM) to be extended in new challenging scenarios. Here are a few concrete proposals:

  1. **Objective framework for metric-based hierarchical clustering and density-based clustering.** Dasgupta's framework [Das16] formalizes similarity-based hierarchical clustering, but many real-world datasets are measured with explicit distance metrics, such as Euclidean or cosine similarity. While distances can sometimes be transformed into similarities via kernels, no universal transformation preserves theoretical guarantees. Similarly, density-based clustering methods (e.g., DBSCAN, OPTICS) are widely used in practice but lack a formal objective function that captures their behavior and quality. These two topics fall into the first category: building theoretical frameworks for learning problems.
  2. **Learning under limited access: sublinear algorithms, oracle-query, and property testing.** The three models are different but related towards the same goal: scalability with approximation guarantee. The first proposal is a follow-up on **Property Testing for Correlation Clustering.** My doctoral work focuses on the standard CC with binary edge labels and our current results are not tight. Beyond closing the gap, we can study *chromatic CC*, where edges have multiple label classes, or *weighted CC*, where edge weights take values in $(0, 1)$. A central question is whether one can design property testers with query complexity $O(1/\text{poly}(1/\varepsilon))$ that distinguish between graphs that admit a good clustering and those that are ε-far. The second proposal is **Learning-Augmented Algorithms.** Learning-augmented oracles arise from the availability of machine learning models that provide predictions. While numerous offline problems have been studied in this framework, there is significant potential in extending these ideas to more general domains, such as online combinatorial optimization and dynamic graph algorithms. Moreover, developing lower bound techniques for learning-augmented algorithms remains an open and promising direction.
  3. **Geometry-grounded learning algorithms.** My doctoral works explore dimension reduction with different geometric assumptions. In fact, many fundamental problems such as clustering, nearest neighbor search exhibit inherent geometric structure, whether in Euclidean, Hyperbolic, or general non-Euclidean spaces. My future research on this line aims to understand the behavior of geometry deviating from Euclidean to highly non-Euclidean, and its impact on these problems. Meanwhile, the goal of geometric-inspired algorithms is to improve efficiency, approximation, and interpretability.

- **Graph Algorithms** Beyond general interest in all graph problems (matching, cut, coloring, flow...) and graph-related structures (spanner, distance oracle, hopset...) in various computation models (streaming, dynamic, distributed...), I highlight a few directions that I intend to pursue.

  1. **Shortest Paths and Beyond.** Our understanding on shortest path remains very limited. Textbook algorithms give $O(m)$ time for SSSP, however the conjecture if APSP requires $\Omega(mn)$ time remains unsolved. Our understanding of shortest paths still lacks a detailed grasp of their combinatorial structures, such as how they overlap and deviate systematically. This is the potential key to downstream algorithms such as differentially private APSD, which still remains a quadratic gap. The same set of questions can be asked for the All k-tuples minimum steiner tree problems, does all-triple require $\Omega(n^4)$ time? How does k impact the hardness? How are these problems connected to the matrix multiplication progress? Resolving these questions is my major pursuit on shortest paths.

     Meanwhile, I welcome new problems motivated from other algorithmic problems or real-world applications. A concrete new problem I want to work on is **Distributed Restricted Shortest Paths.** Shortest path problems are central in Congest model, because they reveal the cost of local-to-global propagation. Restricted SSSP have two edge weights: the weight to compute distance and the cost. The paths cannot have costs higher than a threshold. The goal is to minimize rounds of computation while maintaining a good (approximate) solution.
  2. **Graph Decomposition.** Expander decomposition gathers much interest recently and becomes an essential building block for efficient graph algorithms. It is certainly interesting to push forward its limits, and explore more applications. There are many other decompositions, such as k-core decomposition, sparse-dense decomposition, balanced partition, etc. They are proposed in different context, and I would like to investigate potential relationships among them.

Conclusive Remarks

Thank you for reading thus far. I would like to conclude with a few thoughts that reflect my research philosophy and personal motivations.

- **Happiest moments in research:** Finding out deep connections between two research topics that appear irrelevant from their own context.
- **Which type of researcher?** People say there are two types: researchers come into a field and solve every open problem; or researchers open new fields. I see myself more in the latter.
- **What professions would I pursue if not a researcher?** A philosopher and a musician.

References

[AAD+23] Vikrant Ashvinkumar, Sepehr Assadi, Chengyuan Deng, Jie Gao, and Chen Wang. Evaluating stability in massive social networks: Efficient streaming algorithms for structural balance. *arXiv preprint arXiv:2306.00668*, 2023. 1

[ABD+24] Vikrant Ashvinkumar, Aaron Bernstein, Chengyuan Deng, Jie Gao, and Nicole Wein. Low sensitivity hopsets. *arXiv preprint arXiv:2407.10249*, 2024. 2

[AI08] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122, 2008. 2

[BDG+24] Greg Bodwin, Chengyuan Deng, Jie Gao, Gary Hoppenworth, Jalaj Upadhyay, and Chen Wang. The discrepancy of shortest paths. *arXiv preprint arXiv:2401.15781*, 2024. 2

[CL00] Bernard Chazelle and Alexey Lvov. A trace bound for the hereditary discrepancy. In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 64–69, 2000. 2

[Das16] Sanjoy Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 118–127, 2016. 3

[DCZ+24] Chengyuan Deng, Zhengzhang Chen, Xujiang Zhao, Haoyu Wang, Junxiang Wang, Haifeng Chen, and Jie Gao. Rio-cpd: A riemannian geometric method for correlation-aware online change point detection. *arXiv preprint arXiv:2407.09698*, 2024. 3

[DDJ+25] Chengyuan Deng, Yiqun Duan, Xin Jin, Heng Chang, Yijun Tian, Han Liu, Yichen Wang, Henry Peng Zou, Yijia Xiao, Shenghao Wu, et al. Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas: A survey. *AI and Ethics*, pages 1–27, 2025. 3

[DFW+22] Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. Openfwi: Large-scale multi-structural benchmark datasets for full waveform inversion. *Advances in Neural Information Processing Systems*, 35:6007–6020, 2022. 3

[DGL+24] Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, Hongbin Sun, and Cheng Xin. Neuc-mds: Non-euclidean multidimensional scaling through bilinear forms. *Advances in Neural Information Processing Systems*, 37:121539–121569, 2024. 2

[DGL+25] Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, and Cheng Xin. Johnson lindenstrauss lemma beyond euclidean geometry. *Advances in Neural Information Processing Systems*, 2025. 2

[DGUW23] Chengyuan Deng, Jie Gao, Jalaj Upadhyay, and Chen Wang. Differentially private range query on shortest paths. In *Algorithms and Data Structures Symposium*, pages 340–370. Springer, 2023. 2

[DIIM04] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262, 2004. 2

[FWD+23] Shihang Feng, Hanchen Wang, Chengyuan Deng, Yinan Feng, Yanhua Liu, Min Zhu, Peng Jin, Yinpeng Chen, and Youzuo Lin. e^{FWI}: Multi-parameter benchmark datasets for elastic full waveform inversion of geophysical properties, 2023. 3

[GG05] Ioannis Giotis and Venkatesan Guruswami. Correlation clustering with a fixed number of clusters. *arXiv preprint cs/0504023*, 2005. 1

[IM98] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998. 2

[LZL+23] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*, 2023. 3

[MNP06] Rajeev Motwani, Assaf Naor, and Rina Panigrahi. Lower bounds on locality sensitive hashing. In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 154–157, 2006. 2

[OWZ14] Ryan O'Donnell, Yi Wu, and Yuan Zhou. Optimal lower bounds for locality-sensitive hashing (except when q is tiny). *ACM Transactions on Computation Theory (TOCT)*, 6(1):1–13, 2014. 2