

DNA: A Distribution-and-Aggregation Solution for Spatiotemporal K-function-based Analysis

(Edison) Tsz Nam Chan¹, Bojian Zhu², Dingming Wu¹,
Renchi Yang², Ruisheng Wang¹

¹Shenzhen University

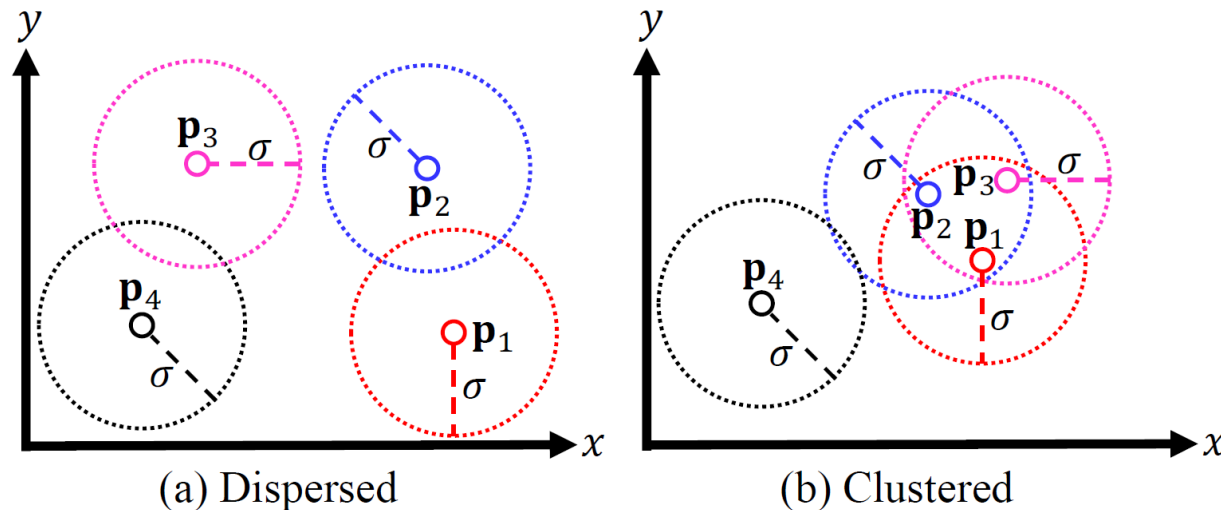
²Hong Kong Baptist University



香港浸會大學
HONG KONG BAPTIST UNIVERSITY

What is K-function?

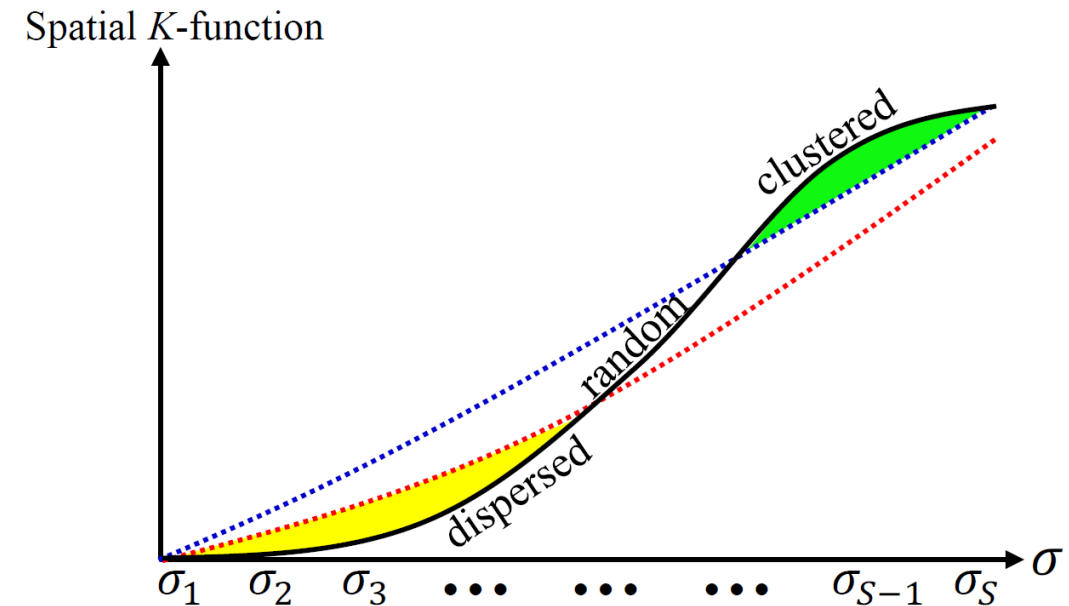
- Count the number of data points that are within the spatial threshold σ from each data point in a dataset.



- If this number is large, the dataset tends to be clustered. Otherwise, the dataset tends to be dispersed.

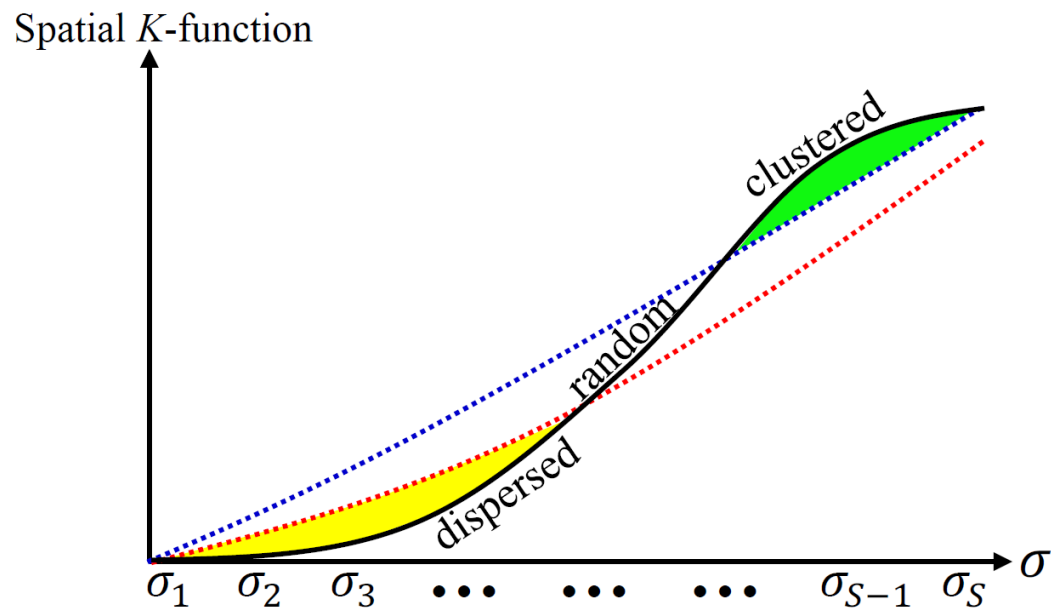
What is K-function Plot?

- Obtain the K-function values (with respect to different thresholds) for the original dataset (black curve).
- Generate L random datasets and obtain the **minimum** and **maximum** K-function values for each threshold (the **red dotted curve** and **blue dotted curve**, respectively).

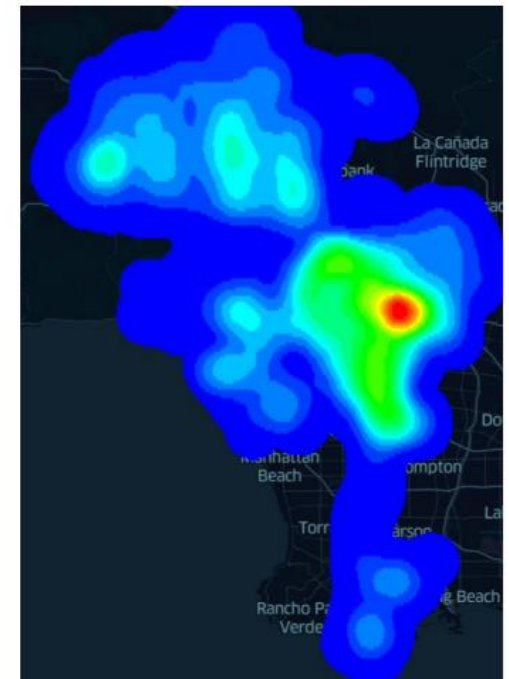


Why K-function Plot?

- Check whether the hotspots/clustering results are meaningful/significant.



Crime events in Los Angeles

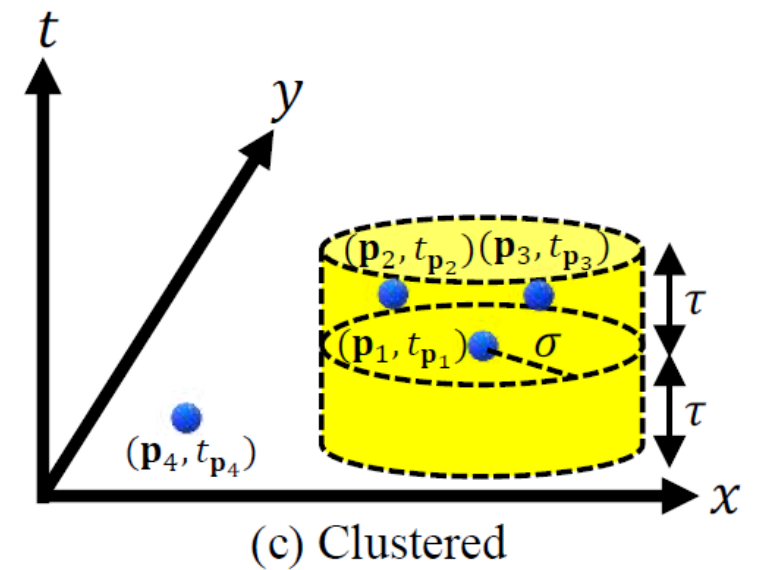
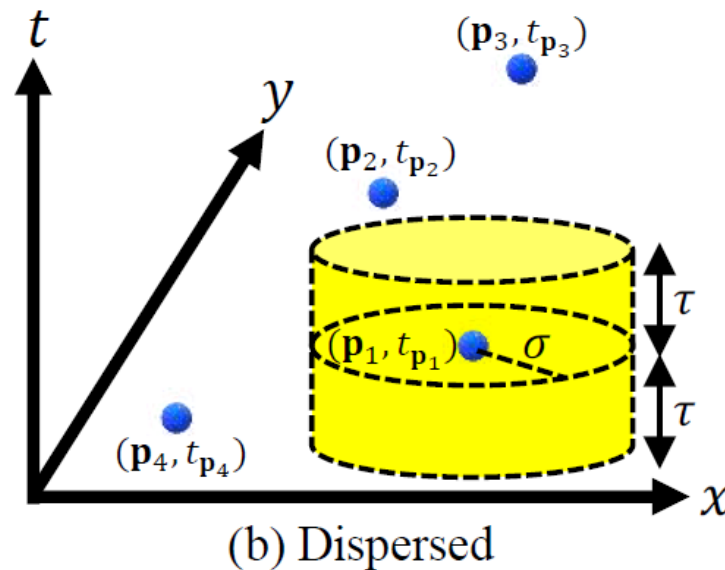
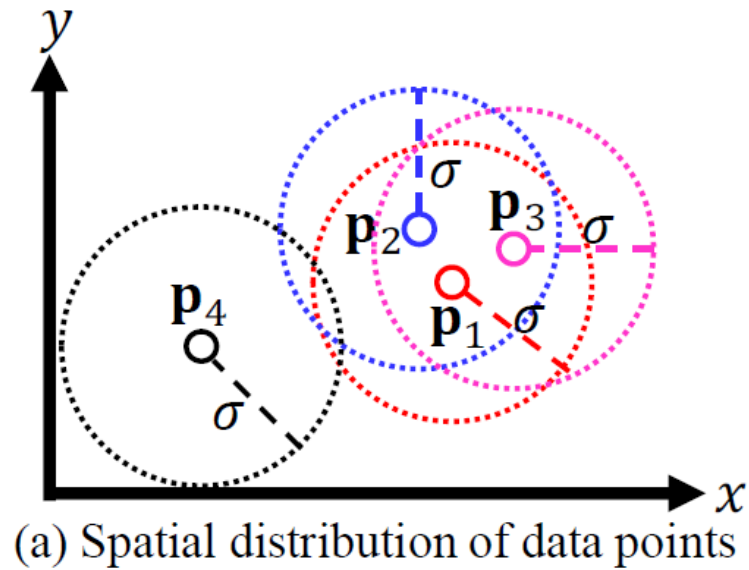


Hotspot map

- Select suitable parameters for some clustering/hotspot detection methods.

K-function Can Be Misleading!

- Does not consider the temporal component of each data point. ☹️

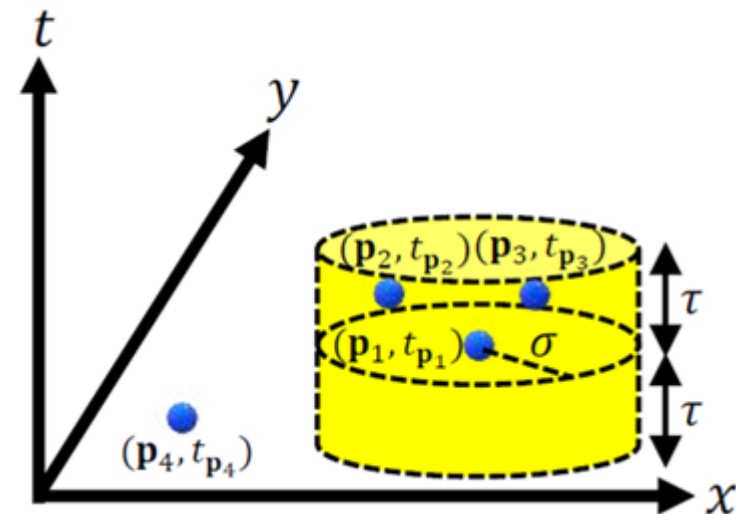


What is Spatiotemporal K-function?

- Consider a set of data points $P = \{(\mathbf{p}_1, t_{\mathbf{p}_1}), (\mathbf{p}_2, t_{\mathbf{p}_2}), \dots, (\mathbf{p}_n, t_{\mathbf{p}_n})\}$ and the spatial threshold σ and the temporal threshold τ . The spatiotemporal K-function is:

$$K_P(\sigma, \tau) = \sum_{(\mathbf{p}_i, t_{\mathbf{p}_i}) \in P} \sum_{\substack{(\mathbf{p}_j, t_{\mathbf{p}_j}) \in P \\ j \neq i}} \mathfrak{I} \left(d(\mathbf{p}_i, \mathbf{p}_j) \leq \sigma, d(t_{\mathbf{p}_i}, t_{\mathbf{p}_j}) \leq \tau \right)$$

where d denotes the Euclidean distance.

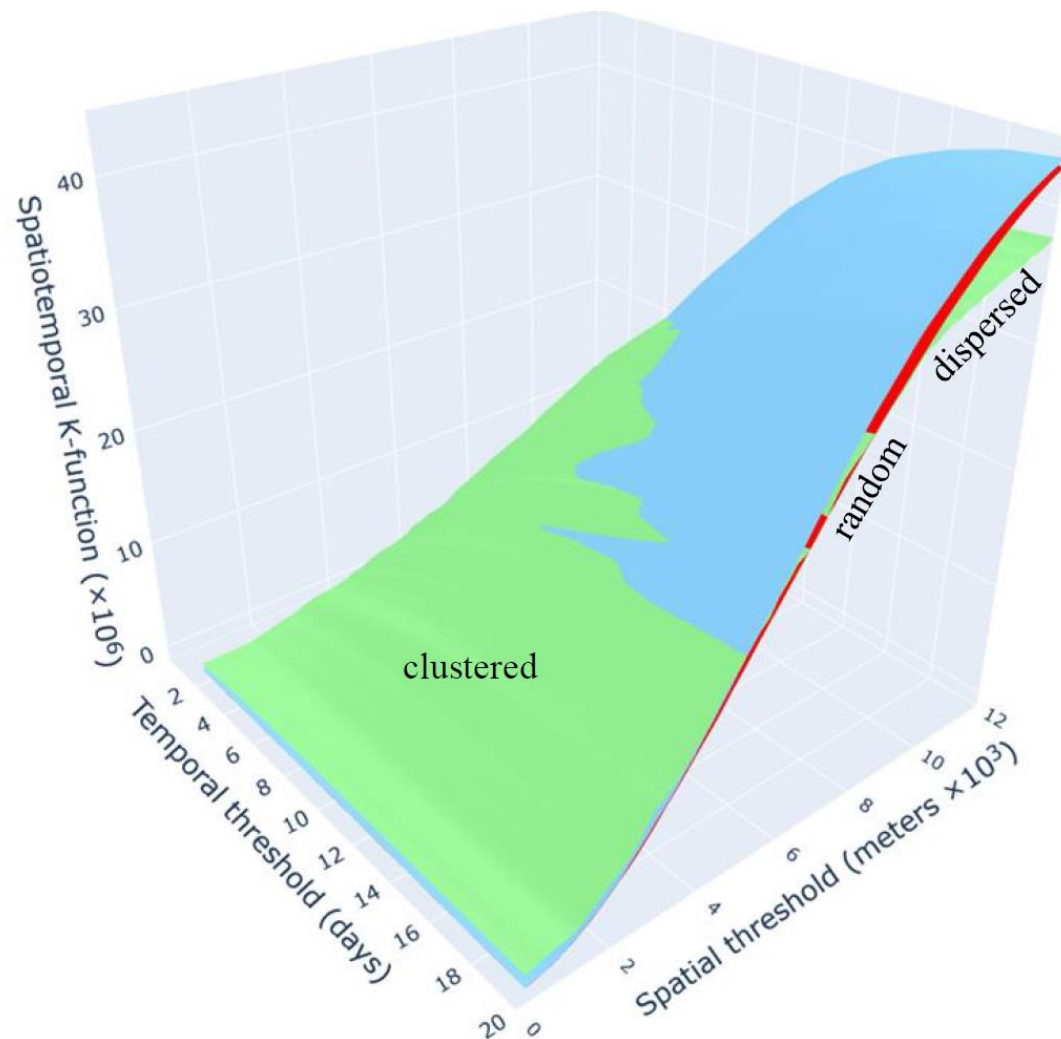


What is Spatiotemporal K-function Plot?

- Compute the spatiotemporal K-functions $K_P(\sigma, \tau)$ for the original dataset P (size n) with respect to S spatial thresholds, $\sigma_1, \sigma_2, \dots, \sigma_S$, and T temporal thresholds, $\tau_1, \tau_2, \dots, \tau_T$ (green plane).
- Generate L random datasets, R_1, R_2, \dots, R_L with the same size n , and compute $\mathcal{L}(\sigma, \tau)$ and $\mathcal{U}(\sigma, \tau)$ for each (σ, τ) -pair.

$$\mathcal{L}(\sigma, \tau) = \min \left(K_{R_1}(\sigma, \tau), K_{R_2}(\sigma, \tau), \dots, K_{R_L}(\sigma, \tau) \right)$$

$$\mathcal{U}(\sigma, \tau) = \max \left(K_{R_1}(\sigma, \tau), K_{R_2}(\sigma, \tau), \dots, K_{R_L}(\sigma, \tau) \right)$$



Spatiotemporal K-function Plot is Slow!

- Time complexity for generating a spatiotemporal K-function plot is $O(LSTn^2)$ ☹️
- Example:
 - Number of random datasets: 4
 - Number of data points: 1,000,000
 - Number of spatial thresholds: 60
 - Number of temporal thresholds: 20
 - Total number of operations: **6000 Trillions** ☹️

Range-Query-based Solution (RQS)

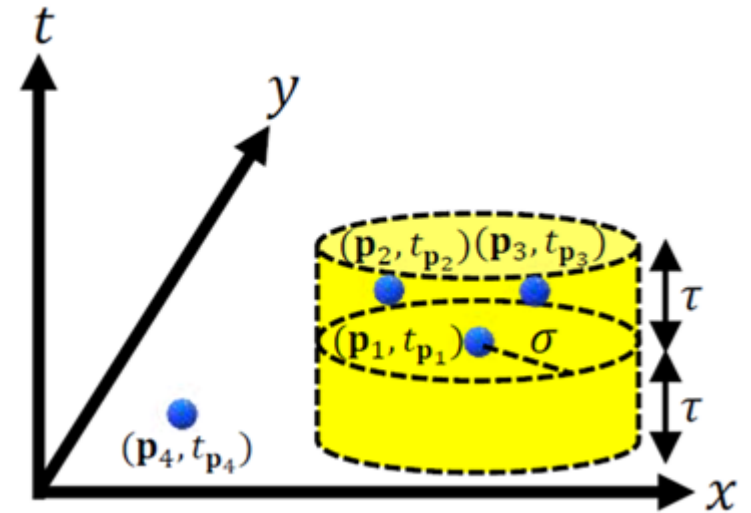
- Find the range query set for each data point $(\mathbf{p}_i, t_{\mathbf{p}_i})$.

$$\mathbb{R}_{\sigma, \tau}(\mathbf{p}_i, t_{\mathbf{p}_i}) = \left\{ (\mathbf{p}_j, t_{\mathbf{p}_j}) \in P : \begin{array}{l} d(\mathbf{p}_i, \mathbf{p}_j) \leq \sigma, d(t_{\mathbf{p}_i}, t_{\mathbf{p}_j}) \leq \tau \\ j \neq i \end{array} \right\}$$

- Compute the Spatiotemporal K-function.

$$K_P(\sigma, \tau) = \sum_{(\mathbf{p}_i, t_{\mathbf{p}_i}) \in P} |\mathbb{R}_{\sigma, \tau}(\mathbf{p}_i, t_{\mathbf{p}_i})|$$

- Many tree structures are available. 😊
- Cannot reduce the time complexity. ☹️



Our Contributions

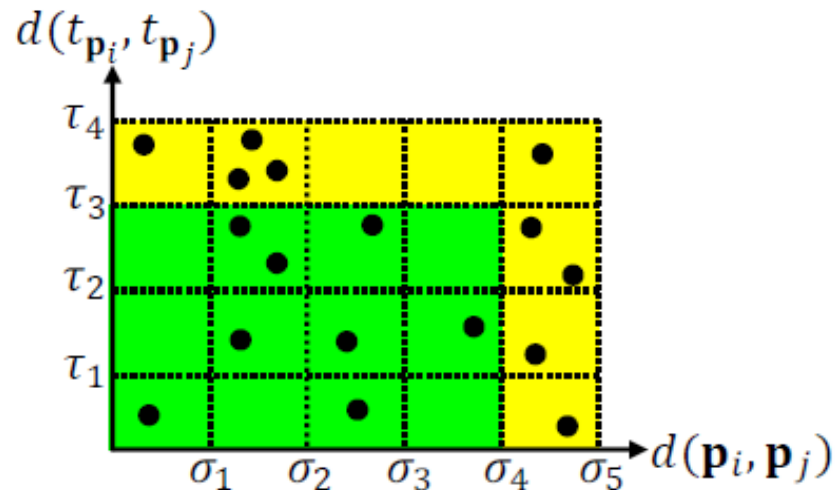
Threshold interval type	Method	Time complexity	Space complexity
Fixed threshold intervals	RQS	$O(LSTn^2)$	$O(nL + ST)$
	DNA	$O(Ln^2 + LSTn)$ (Theorem 1)	$O(nL + ST)$ (Theorem 3)
Multiple threshold intervals	RQS	$O(LSTn^2)$	$O(nL + ST)$
	DNA	$O(Ln^2 \log ST + LSTn)$ (Theorem 2)	$O(nL + ST)$ (Theorem 3)

- Can reduce the worst-case time complexity for generating a spatiotemporal K-function plot, without increasing the space complexity. 😊
- Can achieve 4.58x to 57.42x speedups compared with RQS. 😊

Core Idea: Coverage Property

- The range query set $\mathbb{R}_{\sigma_S, \tau_T}(\mathbf{p}_i, t_{\mathbf{p}_i})$ with the largest spatial threshold σ_S and the largest temporal threshold τ_T can be shared to other range query sets $\mathbb{R}_{\sigma_u, \tau_v}(\mathbf{p}_i, t_{\mathbf{p}_i})$, where $1 \leq u \leq S$ and $1 \leq v \leq S$, i.e.,

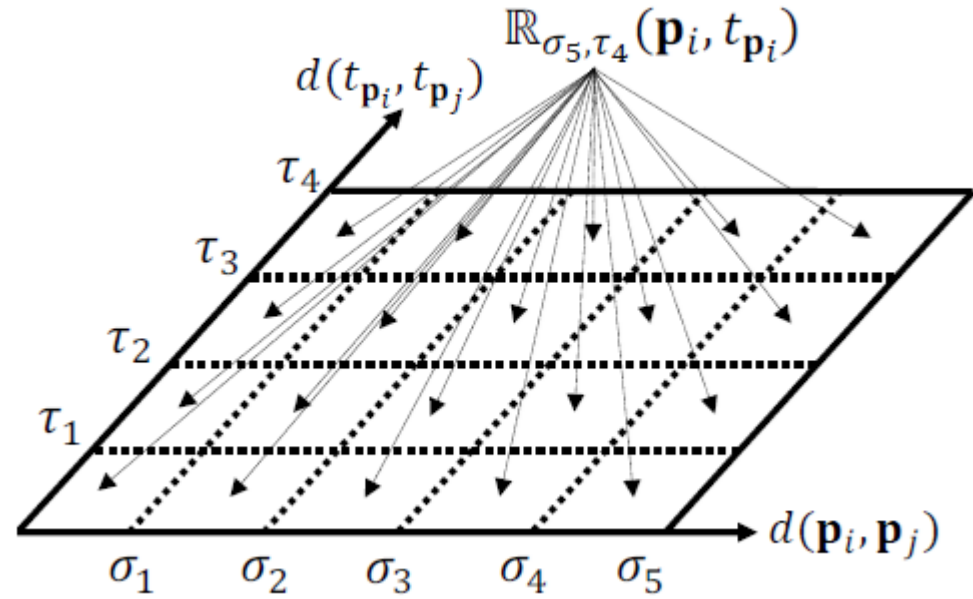
$$\mathbb{R}_{\sigma_u, \tau_v}(\mathbf{p}_i, t_{\mathbf{p}_i}) \subseteq \mathbb{R}_{\sigma_S, \tau_T}(\mathbf{p}_i, t_{\mathbf{p}_i})$$



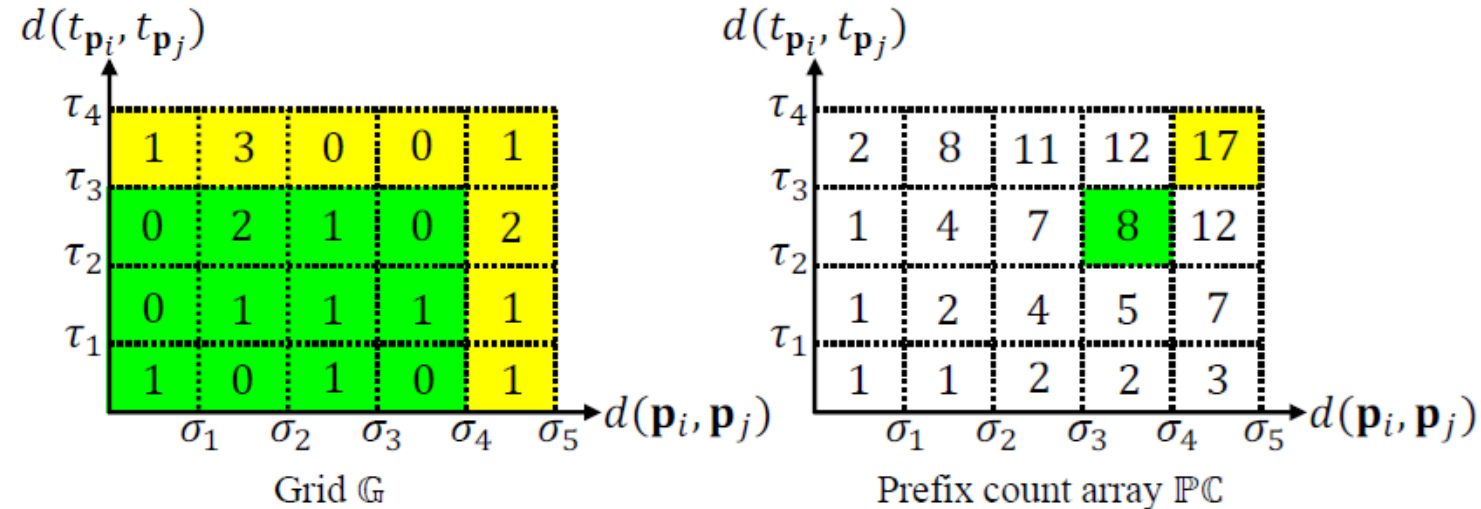
Our Solution: Distribution and Aggregation (DNA)

- Distribution

- Aggregation



$O(n + ST)$ time

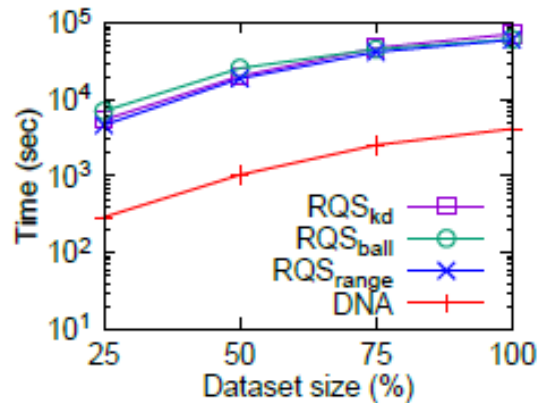


$O(ST)$ time

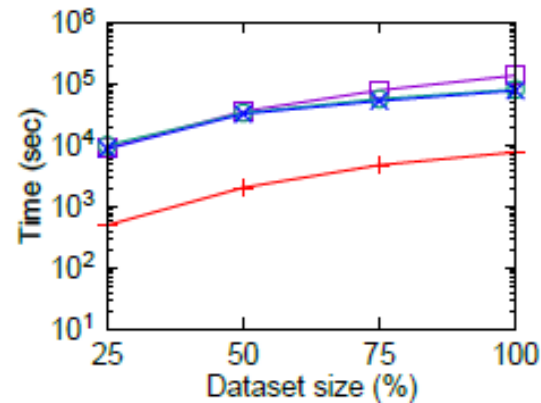
- Total time complexity is $O(Ln^2 + LSTn)$.
- Better than $O(LSTn^2)$. ☺

Experimental Evaluation

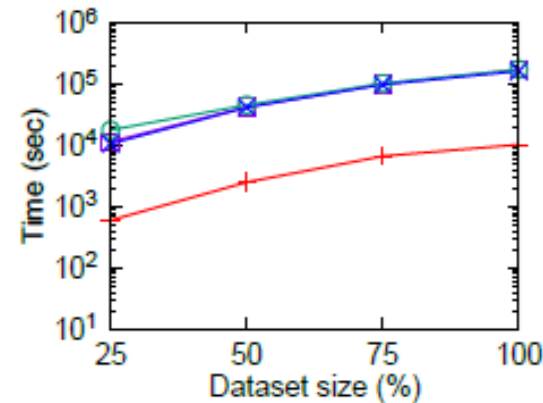
Name	n	Category
Seattle	236,136	Traffic accidents
Atlanta	363,861	Crime events
Los Angeles	559,861	Bicycle mobility
New York	1,541,284	Traffic accidents



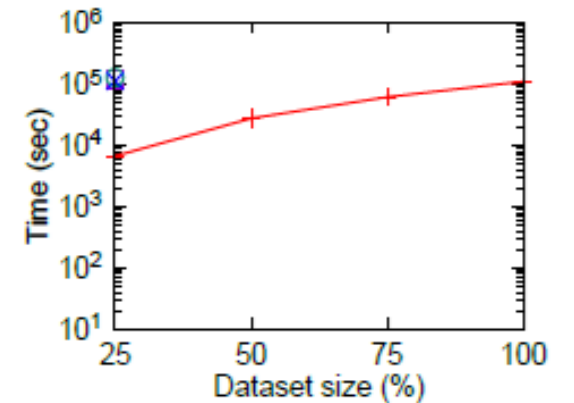
(a) Seattle



(b) Atlanta



(c) Los Angeles



(d) New York

Future Work

- Incorporate DNA into the QGIS/ArcGIS plugin.
- Develop efficient approximate algorithms for supporting spatiotemporal K-function analysis.
- Support other GIS tools, e.g., Moran's I, Getis-Ord General G, and Kriging.
- Leverage modern hardware, e.g., GPU, to further improve the performance of DNA.

Our New Book

Tsz Nam Chan · Dingming Wu
Mastering the Academic Writing Mindset
A Guide to Crafting Computer Science Papers

In the undergraduate study of computer science, a lecturer only teaches some things that are in the literature (most likely in an open access textbook). Those knowledges may have been discovered before in several decades ago. A student is deemed to be good if they have perfectly finished assignments and have prepared well for their examinations. As an example, those students can easily get high grades for all fundamental courses (e.g., programming courses, linear algebra, probability and statistics, data structures, and design and analysis of algorithms) if they have worked extremely hard for the exercises that are provided in those open access textbooks or in class. Therefore, the undergraduate students do not need to have creativity (e.g., establish new knowledges) for obtaining an undergraduate degree. All they need to do is to consolidate their foundation. However, the most critical transition from undergraduate study to postgraduate study is to create new knowledges, which advance the state of the art in the computer science field. Moreover, postgraduate students need to write papers in a logical way (by telling a great story) so that other reviewers can accept them. In order to accomplish these two tasks, students need to change their mindsets for adapting to this new environment. In this book, we discuss this main theme in detail for analyzing the common mistakes that are easily made by new students and show the correct methodology for reading/writing papers. With this methodology, we believe that those students who are dedicated to computer science research can be very productive for publishing top-tier papers.

Except where otherwise noted, this book is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

ISBN 978-981-95-4849-1

springer.com

Chan · Wu

Mastering the Academic Writing Mindset

Tsz Nam Chan · Dingming Wu

Mastering the Academic Writing Mindset

A Guide to Crafting Computer Science Papers

OPEN ACCESS

Springer

<https://link.springer.com/book/10.1007/978-981-95-4850-7>