STARS: A Sampling and Threshold Sharing Solution for Network K-function Analytics

Hongwei Ye
Shenzhen University
yehongwei2023@email.szu.edu.cn

Tsz Nam Chan
Shenzhen University
edisonchan@szu.edu.cn

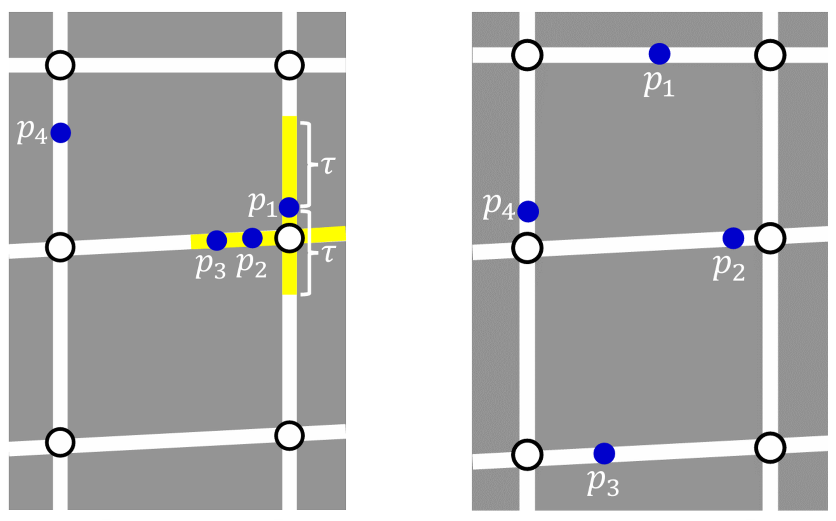
Leong Hou U
University of Macau
ryanlhu@um.edu.mo

Dingming Wu
Shenzhen University
dingming@szu.edu.cn

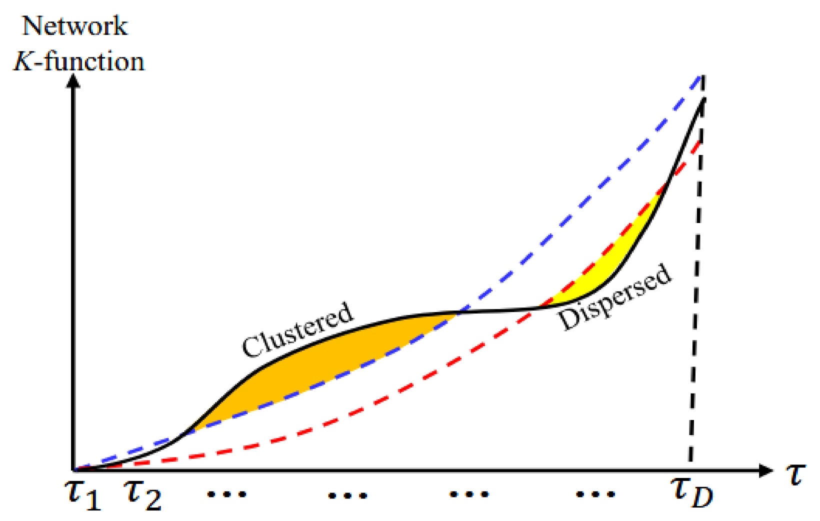
Wei Tu
Ruisheng Wang
Shenzhen University
{tuwei,ruiswang}@szu.edu.cn

Joshua Zhexue Huang
Shenzhen University
zx.huang@szu.edu.cn

What is Network K-function Analytics?



Clustered Dispersed



- **Network K-function** is a fundamental GIS operation for checking whether the clusters/hotspots (discovered by the clustering/hotspot detection algorithms) in the datasets are significant.

$$\underbrace{K_P(\tau)}_{\text{Dataset}} = \frac{1}{n(n-1)} \sum_{p_i\in P} \sum_{\substack{p_j\in P \\ p_j\neq p_i}} \mathbb{I}\big(\overbrace{d_{\underbrace{G}_{\text{Road network}}}(p_i,p_j)}^{\text{Shortest path distance}} \le \underbrace{\tau}_{\text{Threshold}}\big)$$

- Domain experts need to (1) determine D thresholds $\tau_1, \tau_2, \ldots, \tau_D$, (2) generate L random datasets (with the same size n) $R_1, R_2, \ldots, R_L$, and (3) compute the following network K-function values:

(i) $K_P(\tau_d)$
(ii) $\mathcal{L}(\tau_d) = \min\left(K_{R_1}(\tau_d), K_{R_2}(\tau_d), \ldots, K_{R_L}(\tau_d)\right)$
(iii) $U(\tau_d) = \max\left(K_{R_1}(\tau_d), K_{R_2}(\tau_d), \ldots, K_{R_L}(\tau_d)\right)$

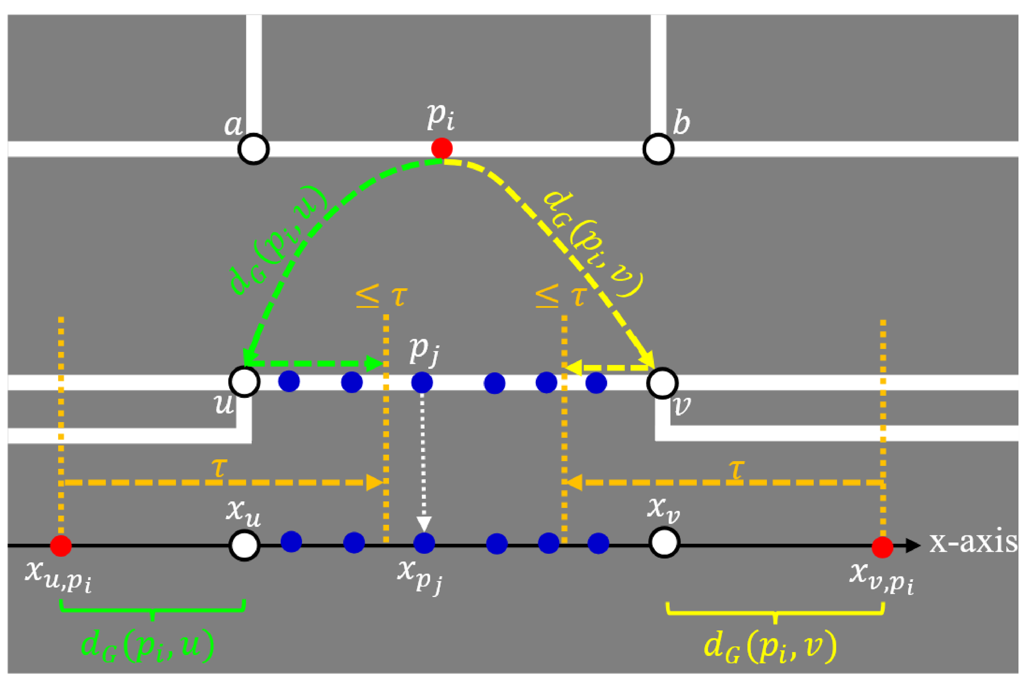
- Existing method are slow!
Example: >**10 hours** on 6.05M events and 100 thresholds

$\boxed{O(|E|T_{sp} + LnD|E| + Ln\log n)}$

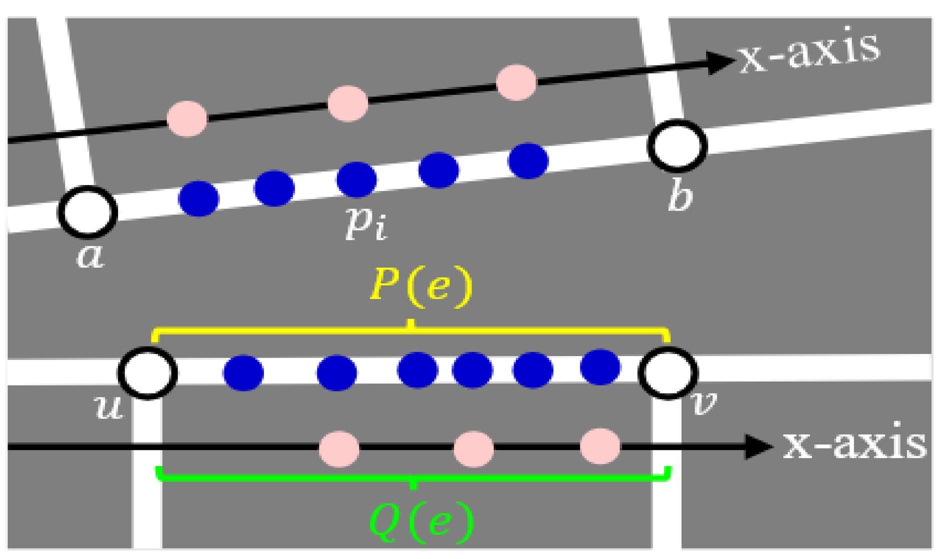
Sampling and Threshold Sharing Solution (STARS)

Sampling

- Reformulate the network K-function as a one-dimensional KDE problem, enabling the use of the optimal Sort Selection (SS) sampling strategy.



① Representative Edge Sampling

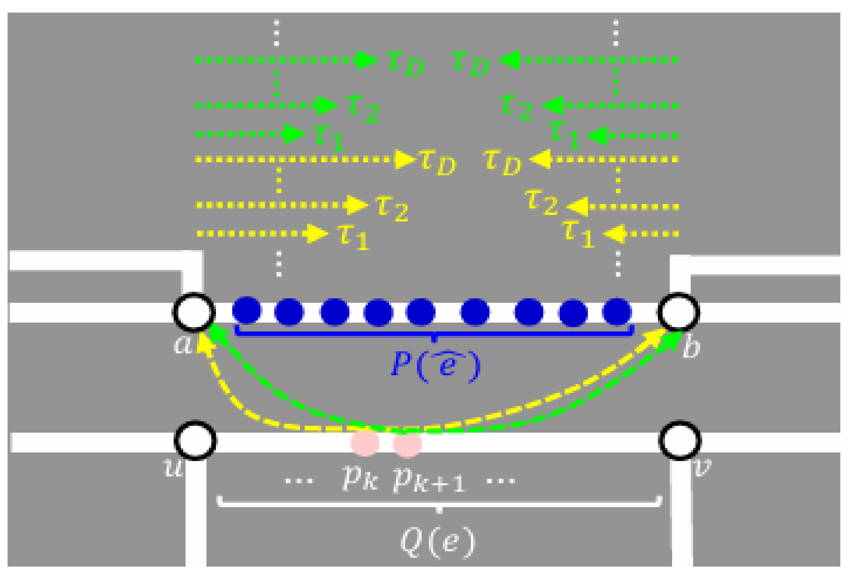


Reduce the evaluated points from $|P(e)|$ to $O(1/\varepsilon)$.

Threshold Sharing

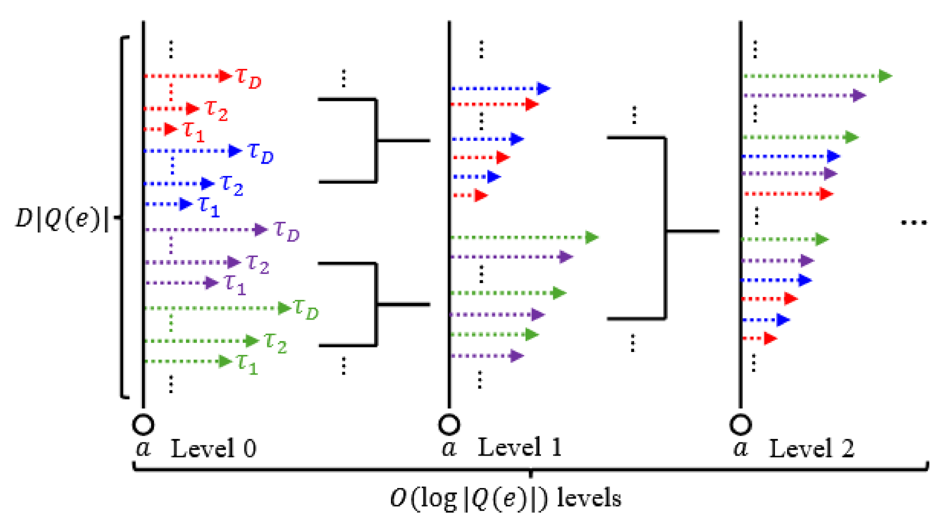
- Merge all threshold boundaries into a single ordered sequence and evaluate all thresholds through one sweep-line pass.

② Generate boundaries



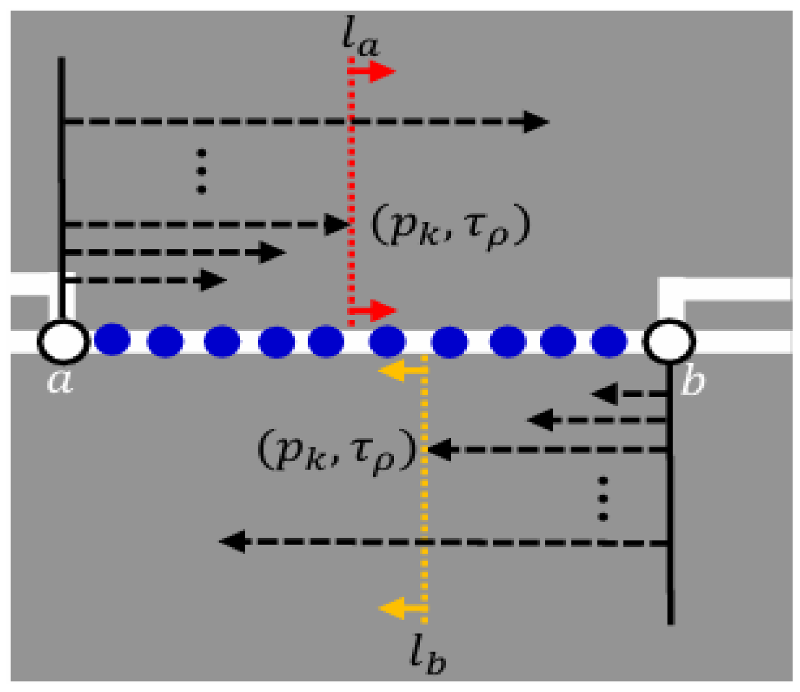
Compute left/right threshold boundaries for every sampled point.

③ Merge & Sort Boundaries



Merge and sort all threshold boundaries into one ordered sequence.

④ Sweep-Line Processing



Sweep once to evaluate all thresholds simultaneously.

Theoretical Guarantees

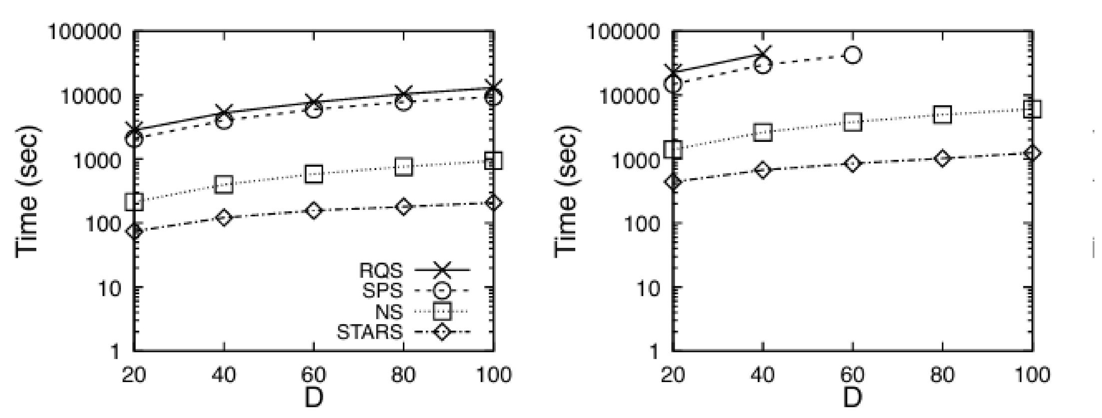
Sampled size
$O(|E|/\varepsilon)$

Approximation Guarantees
$|A_{P,Q}(\tau) - K_P(\tau)| \le \varepsilon$

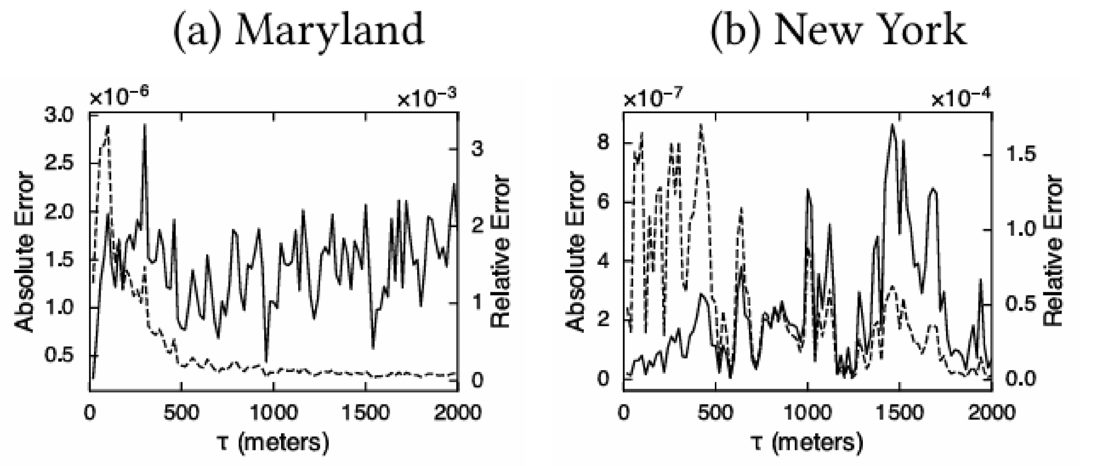
Time Complexity
$O\left(|E|T_{SP} + \frac{LD|E|^2}{\epsilon}\log\left(\frac{1}{\epsilon}\right) + Ln|E| + Ln\log n\right)$

Space Complexity
$O\left(|E| + Ln + \frac{D}{\epsilon} + S_{SP}\right)$

Experimental Results and Case Study

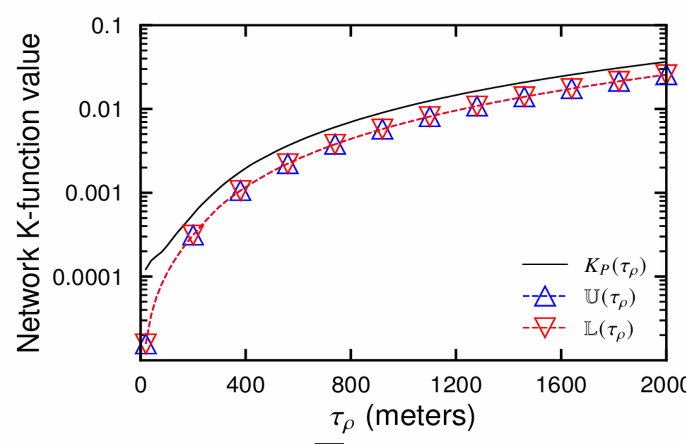


(a) Maryland (b) New York

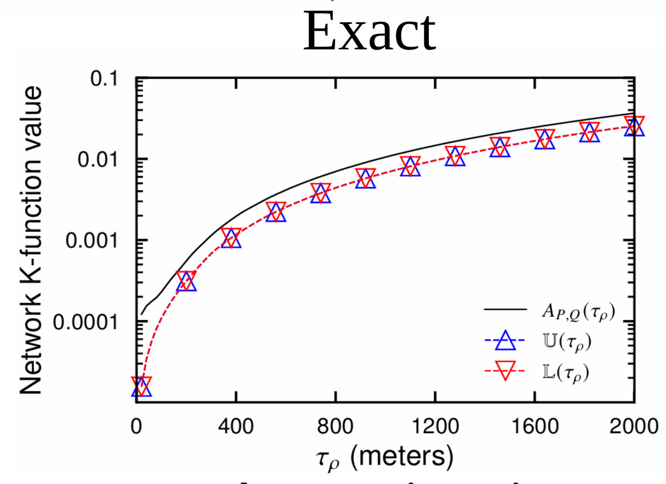


(a) Maryland (b) New York

STARS preserves the analytical conclusions of the exact methods while enabling efficient hotspot analysis



Exact



Approximation