CatchCore: Catching Hierarchical Dense Subtensor

Wenjie Feng, Shenghua Liu, Huawei Shen, and Xueqi Cheng
Institute of Computing Technology ICT, CAS   University of Chinese Academy of Science
fengwenjie@ict.ac.cn, liu.shenghua@gmail.com, {shenhuawei, cxq}@ict.ac.cn

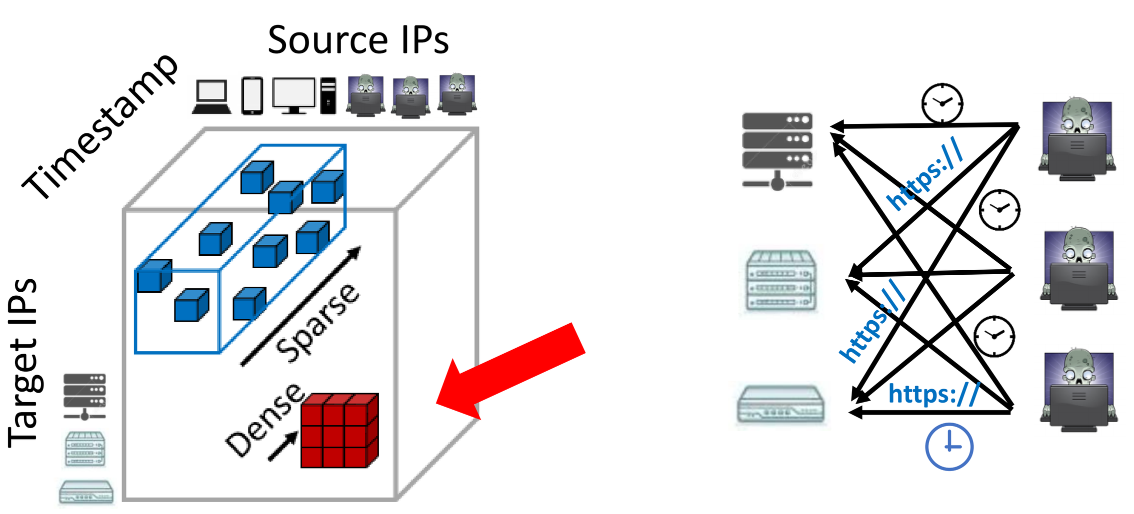
CHINESE ACADEMY OF SCIENCES

Summary

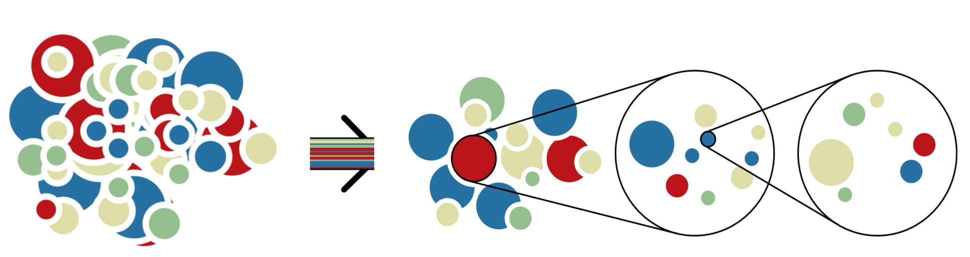
- **Goal: to detect the hierarchical dense subtensors (HDS-tensors)**
- **Previous Work:**
  - Showed that dense subtensors in real-world tensors signal anomalies or fraud
  - Existing related methods assume subtensors are exclusive, and detect flatly and separately
  - Real-world tensors present hierarchical structure properties
- **Proposed Detection method:**
  - **Unified metric:** *(g, h, ϕ)-entry-plenum density measures*
  - **CatchCore:** *gradient based algorithm* detecting *hierarchical dense subtensors*
  - **Quality measure:** *MDL principle to evaluate the detection quality and select optimal parameters*
- **Result:**
  - **Accurate:** CatchCore detect densest subtensor and HDS-tensors with perfect performance
  - **Effective:** our algorithm detect anomalies in various different applications including periodical network intrusion attacks, dense co-authorship research group pattern
  - **Flexible** and **Stable:** our algorithm adapts to many density metrics and is robust with parameters
  - **Provable Scalability:** CatchCore runs in (sub-) linear time with all aspects of tensor

Motivation

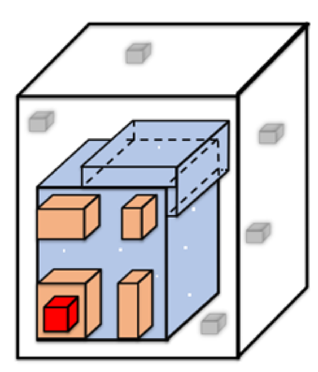
- Dense subtensor in various multi-aspect data (tensor) e.g., the TCP dump for network intrusion, etc.



- Hierarchical structures in real-world data



Community & Core–Periphery

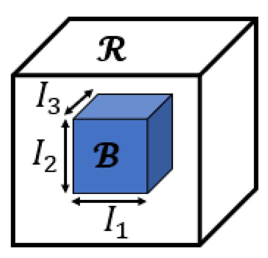


Hierarchical structure

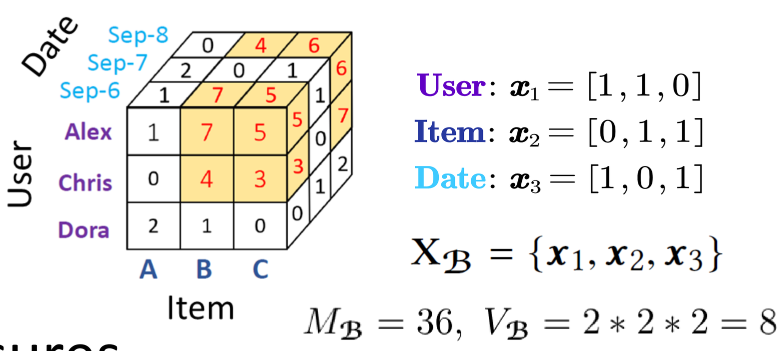
How can we detect hierarchical structures for multi-aspect data?

Terminology

- A subtensor $\boldsymbol{\mathcal{B}}$ in a tensor $\boldsymbol{\mathcal{R}}$
  - Sub-tensor inclusion: $\boldsymbol{\mathcal{B}} \preccurlyeq \boldsymbol{\mathcal{R}}$
  - Mass: $\boldsymbol{M_{\mathcal{B}}}$, Volume: $\boldsymbol{V_{\mathcal{B}}}$, Cardinalities: $\boldsymbol{D_{\mathcal{B}}}$, Density: $\boldsymbol{\rho_{\mathcal{B}}}$



- Indicator vectors paradigm
  - $\mathrm{X}_{\mathcal{B}} = \{\boldsymbol{x}_n \in \{0,1\}^{|\mathcal{R}_n|}; \forall n \in \lfloor N \rceil\}$
  - $M_{\mathcal{B}} = \mathcal{R} \bar{\times} \mathrm{X}_{\mathcal{B}} = \mathcal{R} \times_1 \boldsymbol{x}_1 \cdots \times_N \boldsymbol{x}_N$
  - $\times_n$: *n-mode* tensor-vector product

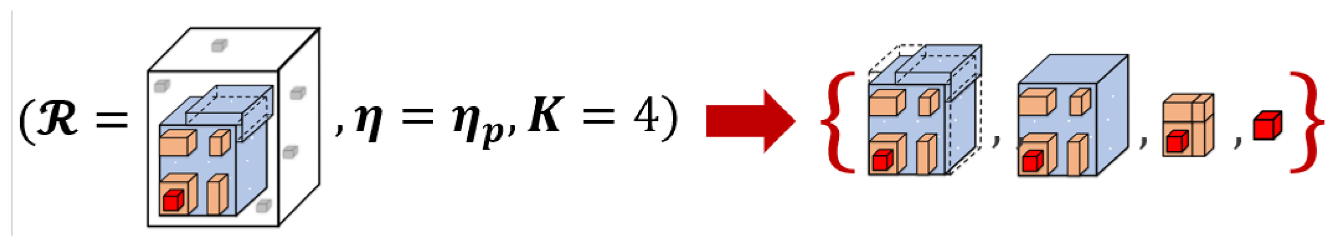


- (g, h, ϕ)-entry-plenum density measures
  - M_{X}, S_{X}: *mass and size (vol. /card.) of subtensor X*
  - *g, h: two increasing functions, ϕ: constant factor*

$$\mathcal{F}_{\phi}(\mathrm{X}) = \begin{cases} 0 & \mathbf{X} = \{\{0\}^{|\mathcal{R}_n|}; \forall n \in \lfloor N \rceil\} \\ g(M_{\mathrm{X}}) - \phi \cdot h(S_{\mathrm{X}}) & \text{otherwise} \end{cases}$$

Proposed Algorithm: CatchCore

- Hierarchical Dense Subtensors Requirements
  - *Density*: significant density difference $\rho(\mathcal{B}^k) \geq \eta\rho(\mathcal{B}^{k-1})\quad \eta > 1$
  - *Structure*: multi-layer cores $\mathcal{B}^k \preccurlyeq \mathcal{B}^{k-1}$, *i.e.*, $\mathcal{B}_n^k \subseteq \mathcal{B}_n^{k-1}, \forall n \in \lfloor N \rceil$
- **Given:**
  $\boldsymbol{\mathcal{R}}$: *N*-way tensor; η: density ratio; K: the max #hierarchies
- **Find:**
  r ($\leq K$) significant HDS tensors $\{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^r\}$



- **Densest subtensor (block)** detection formulation

$$\max_{\mathrm{X}:\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}} \mathcal{F}(\mathrm{X}) = (1+p)\mathcal{R} \bar{\times} \mathrm{X} - p \prod_{\boldsymbol{x}_n \in \mathrm{X}} ||\boldsymbol{x}_n||_1$$
$$\text{subject to } \boldsymbol{x}_n \in \{0,1\}^{|\mathcal{R}_n|},\ \forall n = 1, \cdots, N.$$

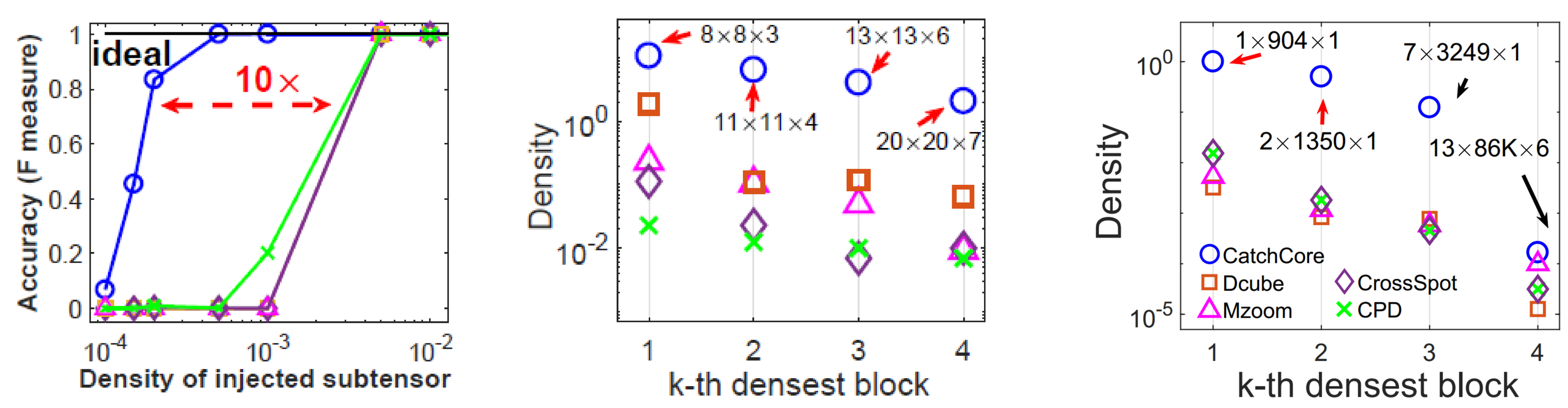
- **HDS-tensors** detection Alg.
  - **Gradient based Optimization** Algorithm
  - **Alternative Gauss-Seidel** for updating

$$\max_{\mathbf{X}^1,\ldots,\mathbf{X}^K} \sum_{k=1}^{K} \mathcal{F}(\mathbf{X}^k)$$
$$\text{subject to } \rho_{\mathbf{X}^{h+1}} \geq \eta \rho_{\mathbf{X}^h}$$
$$\mathbf{X}_{(h+1,n,\cdot)} \leq \mathbf{X}_{(h,n,\cdot)} \leq \mathbf{X}_{(h-1,n,\cdot)}$$
$$\forall h = 1, \cdots, K; \forall n = 1, \cdots, N$$

- Using MDL to evaluate the quality of detection result of given parameter configuration: $M^* = \mathit{argmin}\ L(M) + L(D|M)$

Experiment Results

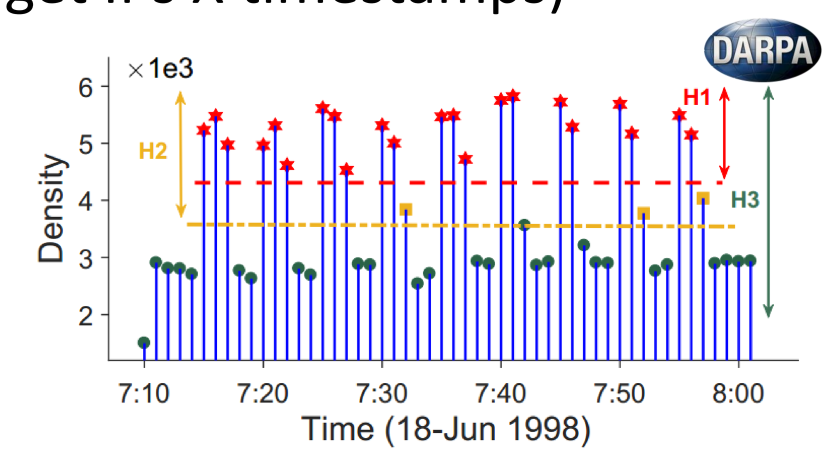
- **Q1 Accuracy:** detect the densest block / HDS-tensors



Inject dense block   dblp   YouTube

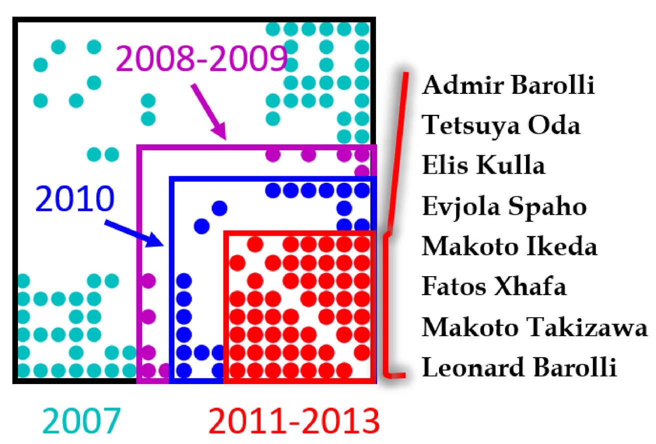
- **Q2 Effectiveness:** patterns detected in real-world tensors
  - (1) TCP dump: (source IPs X target IPs X timestamps)

H	Shape	Ratio
1	$1\times1\times97$	100%
2	$1\times1\times100$	100%
3	$1\times1\times274$	100%
4	$15\times5\times24.7K$	87.0%
5	$171\times15\times29.2K$	85.4%



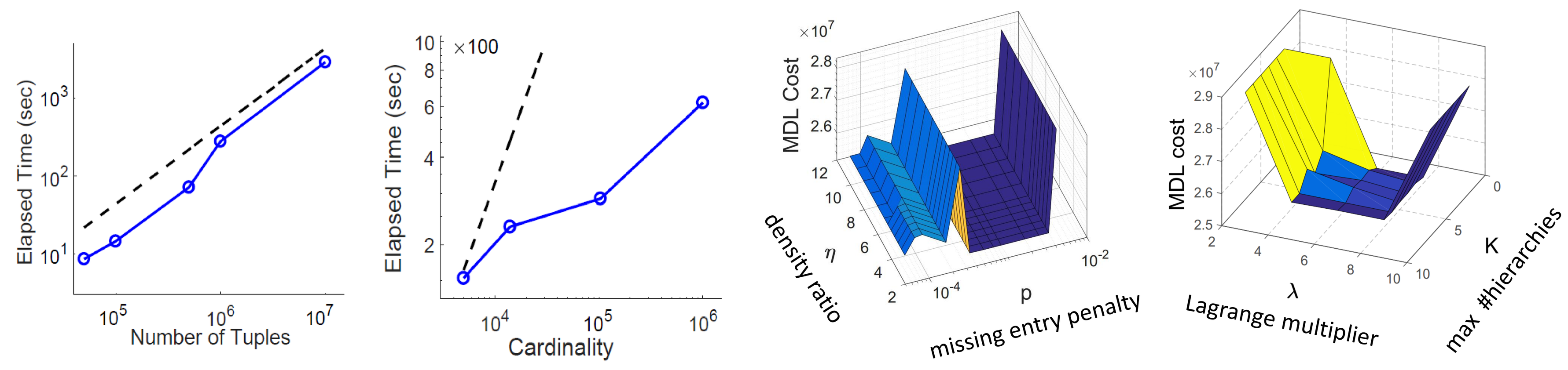
Top 5 HDS-tensors & attack ratio   Periodical Neptune attack

  - (2) DBLP co-authorship: (authors X authors X years)



Top 4 HDS-tensors   Research group

- **Q3 Scalability & Robustness:** *CatchCore is linear in all aspects of input, achieves the optimal results for wide parameters range*



(sub-) Linearly scale up   Insensitive to parameters