M-Zoom: Fast Dense-Block Detection in Tensors with Quality Guarantees - Supplementary Document

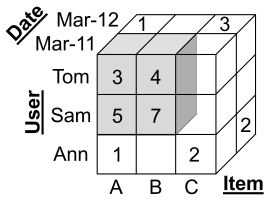
Abstract. In this supplementary document, we provide additional examples, proofs, data descriptions, and experimental results, all of which supplement the main paper [2].

A Example Dataset

Example 1 (Purchase History). Let $\mathcal{R} = Purchase(user, item, date, count)$, depicted in Figure 6a. Each tuple (u, i, d, c) in $\mathcal{R}$ indicates that user u purchased c units of item i on date d. The first three attributes, $A_1 = user$, $A_2 = item$, and $A_3 = date$, are dimension attributes, and the other one, $X = count$, is the measure attribute. Let $\mathcal{B}_1 = \{$'Tom', 'Sam'$\}$, $\mathcal{B}_2 = \{$'A','B'$\}$, and $\mathcal{B}_3 = \{$Mar-11$\}$. Then, $\mathcal{B}$ is the set of tuples regarding the purchases by 'Tom' or 'Sam' on 'A' or 'B' on Mar-11, and its mass $M_{\mathcal{B}} = 19$, the total units sold by such purchases. Likewise, $M_{\mathcal{B}(\text{'Tom'})} = Mass(\mathcal{B}(\text{'Tom'})) = 7$, the total units of 'A' or 'B' purchased by exactly 'Tom' on Mar-11. In the tensor representation, $\mathcal{B}$ composes a subtensor in $\mathcal{R}$, as depicted in Figure 6b.

User	Item	Date	Count
Tom	A	Mar-11	3
Tom	B	Mar-11	4
Sam	A	Mar-11	5
Sam	B	Mar-11	7
Ann	C	Mar-12	2
⋮	⋮	⋮	⋮

(a) Relation $\mathcal{R}$



(b) Tensor Representation of $\mathcal{R}$

Fig. 6: Pictorial description of Example 1. (a) Relation $\mathcal{R}$ (*Purchase*). The shaded tuples compose block $\mathcal{B}$. (b) Tensor representation of $\mathcal{R}$. In the tensor representation, $\mathcal{B}$ forms a subtensor of $\mathcal{R}$.

B Proof of Theorem 3

Proof. Algorithm 2 in main paper requires $O(N|\mathcal{R}|)$ space for $\mathcal{R}$ and $\mathcal{B}$; and $O(\sum_{n=1}^{N} |\mathcal{R}_n|)$ space for min-heaps and the order by which attribute values are removed, as explained in Section 3.2 of the main paper. The sum is $O(N|\mathcal{R}| + \sum_{n=1}^{N} |\mathcal{R}_n|) = O(N|\mathcal{R}|)$ since $|\mathcal{R}_n| \leq |\mathcal{R}|$, $\forall n \in [N]$. Since Algorithm 1 requires additional $O(kN|\mathcal{R}|)$ for storing k blocks it finds, its space complexity is $O(N|\mathcal{R}| + kN|\mathcal{R}|) = O(kN|\mathcal{R}|)$. □

C Proof of Theorem 5

Let $\mathcal{B}^{(r)}$ be the relation $\mathcal{B}$ at the beginning of the r-th iteration of Algorithm 2, and $a_i^{(r)} \in \mathcal{B}_i^{(r)}$ be the attribute value removed in the same iteration, as in the main paper.

Lemma 4. *For any $\alpha \in [0,1]$, there exists a block $\mathcal{B}'$ satisfying $\forall a_i \in \bigcup_{n=1}^{N} \mathcal{B}'_n$, $Mass(\mathcal{B}'(a_i)) \geq \alpha\rho_{ari}(\mathcal{R},\mathcal{R})/N$ and $Mass(\mathcal{B}') \geq (1-\alpha)Mass(\mathcal{R})$.*

Proof. Let s be the first iteration in Algorithm **??** where $Mass(\mathcal{B}^{(s)}(a_i^{(s)})) \geq \alpha\rho_{ari}(\mathcal{R},\mathcal{R})/N$. Such s exists, otherwise

$$Mass(\mathcal{R}) = \sum_{r=1}^{Size(\mathcal{R})} Mass(\mathcal{B}^{(r)}(a_i^{(r)})) < \frac{\alpha\rho_{ari}(\mathcal{R},\mathcal{R})}{N} Size(\mathcal{R}) = \alpha Mass(\mathcal{R}),$$

which is a contradiction. Then,

$$\begin{aligned} Mass(\mathcal{R}) &= \sum_{r=1}^{Size(\mathcal{R})} Mass(\mathcal{B}^{(r)}(a_i^{(r)})) \\ &= \sum_{r=1}^{s-1} Mass(\mathcal{B}^{(r)}(a_i^{(r)})) + \sum_{r=s}^{Size(\mathcal{R})} Mass(\mathcal{B}^{(r)}(a_i^{(r)})) \\ &\leq s\alpha\rho_{ari}(\mathcal{R},\mathcal{R})/N + Mass(\mathcal{B}^{(s)}) \\ &\leq \alpha Size(\mathcal{R})\rho_{ari}(\mathcal{R},\mathcal{R})/N + Mass(\mathcal{B}^{(s)}) = \alpha Mass(\mathcal{R}) + Mass(\mathcal{B}^{(s)}) \end{aligned}$$

Thus, $Mass(\mathcal{B}^{(s)}) \geq (1-\alpha)Mass(\mathcal{R})$. □

Proof of Theorem 5

Proof. Let $\alpha = N/(N+1)$. By Lemma 4 (in the main paper), there exists a block $\bar{\mathcal{B}} \subset \mathcal{B}^*$ satisfying $\forall a_j \in \bigcup_{n=1}^{N} \bar{\mathcal{B}}_n$, $Mass(\bar{\mathcal{B}}(a_j)) \geq \alpha\rho_{ari}(\mathcal{B}^*,\mathcal{B}^*)/N = \alpha\rho_{ari}(\mathcal{B}^*,\mathcal{R})/N$ and $Mass(\bar{\mathcal{B}}) \geq (1-\alpha)Mass(\mathcal{B}^*)$. Let s be the first iteration of Algorithm 2 (in the main paper) where $a_i^{(s)} \in \bigcup_{n=1}^{N} \bar{\mathcal{B}}_n$. By Lemma 3 (in Appendix C of the main paper) and $\mathcal{B}^{(s)} \supset \bar{\mathcal{B}}$, $\forall a_j \in \bigcup_{n=1}^{N} \mathcal{B}_n^{(s)}$, $Mass(\mathcal{B}^{(s)}(a_j)) \geq \alpha\rho_{ari}(\mathcal{B}^*,\mathcal{R})/N$ and $Mass(\mathcal{B}^{(s)}) \geq (1-\alpha)Mass(\mathcal{B}^*)$. If $Size(\mathcal{B}^{(s)}) \geq S_{min}$,

$$\begin{aligned} &\rho_{ari}(\mathcal{B}',\mathcal{R}) \geq \rho_{ari}(\mathcal{B}^{(s)},\mathcal{R}) = \frac{Mass(\mathcal{B}^{(s)})}{Size(\mathcal{B}^{(s)})/N} \\ &= \frac{\sum_{a_j \in \bigcup_{n=1}^{N} \mathcal{B}_n^{(s)}} Mass(\mathcal{B}^{(s)}(a_j))}{Size(\mathcal{B}^{(s)})} \geq \frac{\alpha Size(\mathcal{B}^{(s)})\rho_{ari}(\mathcal{B}^*,\mathcal{R})/N}{Size(\mathcal{B}^{(s)})} = \frac{\rho_{ari}(\mathcal{B}^*,\mathcal{R})}{N+1}. \end{aligned}$$

If $Size(\mathcal{B}^{(s)}) < S_{min}$, we consider $\mathcal{B}^{(q)}$ where $Size(\mathcal{B}^{(q)}) = S_{min}$ and thus $q < s$. Then,

$$\rho_{ari}(\mathcal{B}', \mathcal{R}) \geq \rho_{ari}(\mathcal{B}^{(q)}, \mathcal{R}) = \frac{Mass(\mathcal{B}^{(q)})}{Size(\mathcal{B}^{(q)})/N} \geq \frac{Mass(\mathcal{B}^{(s)})}{Size(\mathcal{B}^{(q)})/N}$$
$$\geq \frac{(1-\alpha)Mass(\mathcal{B}^*)}{Size(\mathcal{B}^{(q)})/N} = \frac{Mass(\mathcal{B}^*)/(N+1)}{S_{min}/N} \geq \frac{Mass(\mathcal{B}^*)/(N+1)}{Size(\mathcal{B}^*)/N} = \frac{\rho_{ari}(\mathcal{B}^*, \mathcal{R})}{N+1}.$$

Hence, regardless of $Size(\mathcal{B}^{(s)})$, $\rho_{ari}(\mathcal{B}', \mathcal{R}) \geq \rho_{ari}(\mathcal{B}^*, \mathcal{R})/(N+1)$. □

We remark that the above proof of Theorem 5 is a multi-dimensional generalization the proof of Theorem 1 in [1].

D Description of AirForce Dataset

The descriptions of the attributes in AirForce Dataset are as follows:

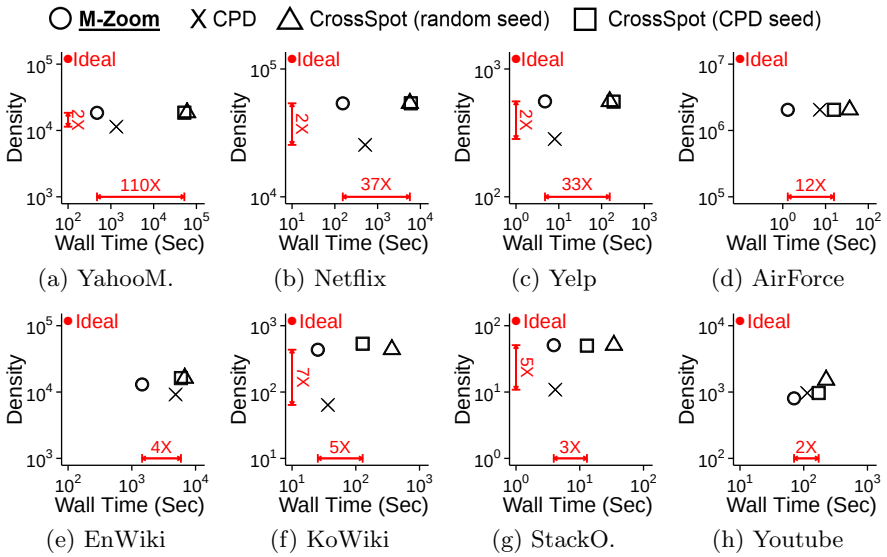
- *protocol* (A_1): type of protocol (e.g. tcp, udp, etc.)
- *service* (A_2): type of network service on destination (e.g., http, telnet, etc)
- *src_bytes* (A_3): amount of data bytes from source to destination
- *dst_bytes* (A_4): amount of data bytes from destination to source
- *flag* (A_5): normal or error status of each connection
- *host_count* (A_6): number of connections to the same host in the past two seconds
- *srv_count* (A_7): number of connections to the same service in the past two seconds
- *#connections* (X): number of connections with the corresponding dimension attribute values.

E Additional Experiments

E.1 Running Time and Accuracy of M-Zoom with Different Density Measures

As in Section 4.2 of the main paper, we compare the speed of different methods and the densities of the blocks found by the methods in real-world datasets. In this section, however, ρ_{geo} and ρ_{susp} (Definitions 2 and 3 in the main paper) were used as the density metrics instead of ρ_{ari}. For each metric, we measured time taken to find three blocks and the maximum density among the three blocks.

Figure 7 shows the result when ρ_{geo} was used as the density metric. M-Zoom provided a significantly better trade-off between speed and accuracy than the other methods in most datasets. For example, in YahooM. Dataset, M-Zoom was 110 times faster than CrossSpot but still found blocks with similar densities. In addition, in the same dataset, M-Zoom detected blocks twice as dense as those detected by CPD and was still 2.8 times faster than CPD. The similar results were obtained using ρ_{susp} as the density measure, as seen in Figure 8.



(a) YahooM. (b) Netflix (c) Yelp (d) AirForce

(e) EnWiki (f) KoWiki (g) StackO. (h) Youtube

Fig. 7: M-Zoom outperforms its competitors when ρ_{geo} is used as the density measure. In each plot, points represent the speed of different methods and the highest density (ρ_{geo}) of three blocks found by the methods. Upper-left region indicates better performance. In most datasets, M-Zoom is the only method that achieves both speed and accuracy.

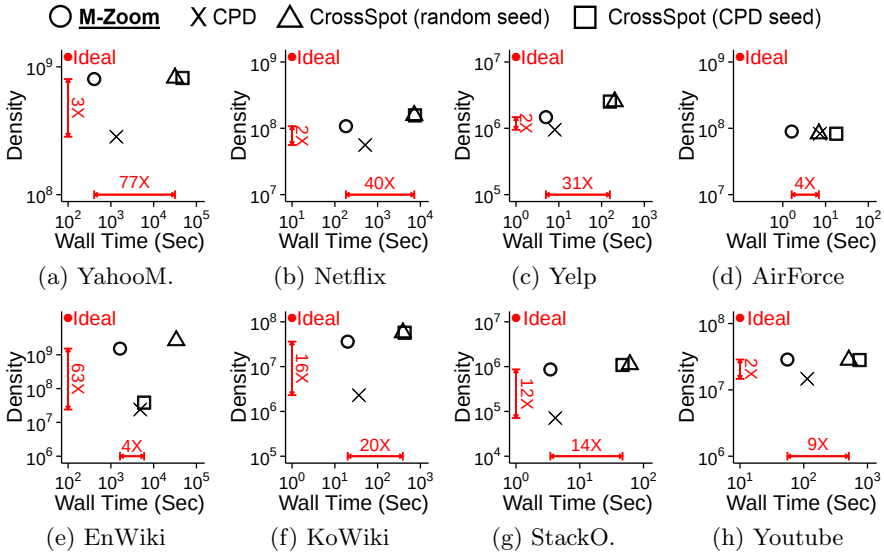
E.2 Diversity of Blocks Found by M-Zoom with Different Density Measures

As in Section 4.3 of the main paper, we compare the diversity of dense blocks found by each method. In this section, however, ρ_{geo} and ρ_{susp} (Definitions 2 and 3 in the main paper) were used as the density metrics instead of ρ_{ari}. For each density metric, the diversity of the blocks found by each method was measured in the same way as in the main paper.

As seen in Figure 9, in all real-world datasets, M-Zoom and CPD successfully detected distinct dense blocks regardless of the density measure used. However, in many datasets, CrossSpot found the same block repeatedly or blocks with slight difference, even when it started from different seed blocks.

References

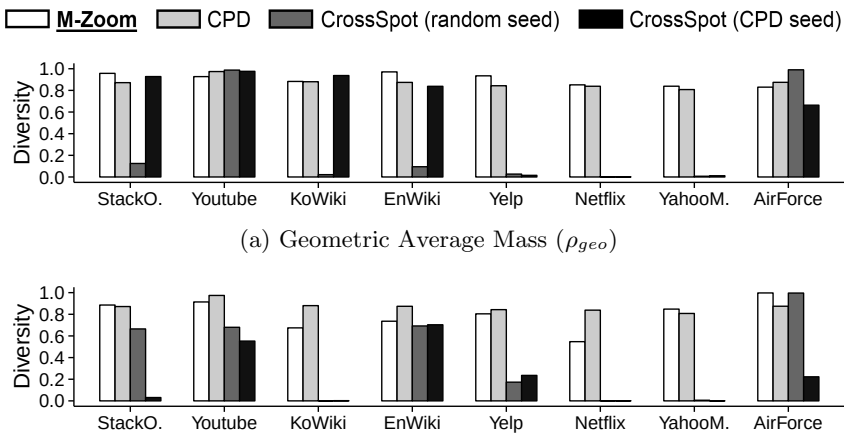
1. Andersen, R., Chellapilla, K.: Finding dense subgraphs with size bounds. In: WAW (2009)
2. Shin, K., Hooi, B., Faloutsos, C.: M-zoom: Fast dense-block detection in tensors with quality guarantees. In: PKDD (2016)



(a) YahooM. (b) Netflix (c) Yelp (d) AirForce

(e) EnWiki (f) KoWiki (g) StackO. (h) Youtube

Fig. 8: M-Zoom outperforms its competitors when ρ_{susp} is used as the density measure. In each plot, points represent the speed of different methods and the highest density (ρ_{susp}) of three blocks found by the methods. Upper-left region indicates better performance. In most datasets, M-Zoom is the only method that achieves both speed and accuracy.



(a) Geometric Average Mass (ρ_{geo})

(b) Suspiciousness (ρ_{susp})

Fig. 9: M-Zoom detects many different dense blocks regardless of the density measure used. The dense blocks found by M-Zoom and CPD have high diversity in all datasets, while the dense blocks found by CrossSpot are almost same in many datasets.