B INTERPRETABILITY OF COMPONENTS

If each component has a specific pattern rather than a mixture of multiple patterns, the component is easy to distinguishable from the others, making the model more interpretable. In our experiment, we use a DBLP dataset to evaluate latent patterns based on labels, consisting of (author, paper, conference) tuples. Each entity of modes is labeled according to the area of study (DM, AI, DB, IR), which is also used in Section 4.3.1. To evaluate this interpretability, we leverages two metrics:

- **Entropy of the top-K label distribution per component.** This metric evaluates whether each component per mode represents a distinct latent pattern. A lower entropy indicates less uncertainty about the labels in each column (component), signifying a higher occurrence of certain labels and making each component easily distinguishable.
- **Distance between the top-K label distributions of the rth components.** This metric evaluates whether how well the rth components across the modes cluster together based on specific labels. The lower the distance is, the better components clusters.

To find the top-K entities for each column of each factor matrix, we sort the factor values in descending order. Then, we create a top-K label distribution using the labels of the top-K entities. We will explain in detail how to measure the entropy and distance in the following paragraph.

$L = \{L_1, \ldots, L_m, \ldots L_M\}$ is a set of labels. If this label is binary, then $M = 2$ and $|L| = 2$. $L^{(n)} = \{(i_n, l_{i_n}) | 1 \leq i_n \leq I_n, l_{i_n} \in L\}$ is a set of indices i_n paired with their corresponding labels l_{i_n}, for $n = 1, \ldots, N$. The average entropy (AE) of top-K label distribution is defined as follows.

$$AE = \frac{1}{N \times R} \sum_{n=1}^{N} \sum_{r=1}^{R} H(\mathbf{p}_r^{(n)}) \text{ where } H(\mathbf{p}_r^{(n)}) = -\sum_{m=1}^{M} \mathbf{p}_r^{(n)}(m) \log \mathbf{p}_r^{(n)}(m). \tag{11}$$

Here, $\mathbf{p}_r^{(n)} \in \mathbb{R}^{M \times 1}$ denotes the top-K label distribution for rth column in the nth factor matrix, which is defined as follows.

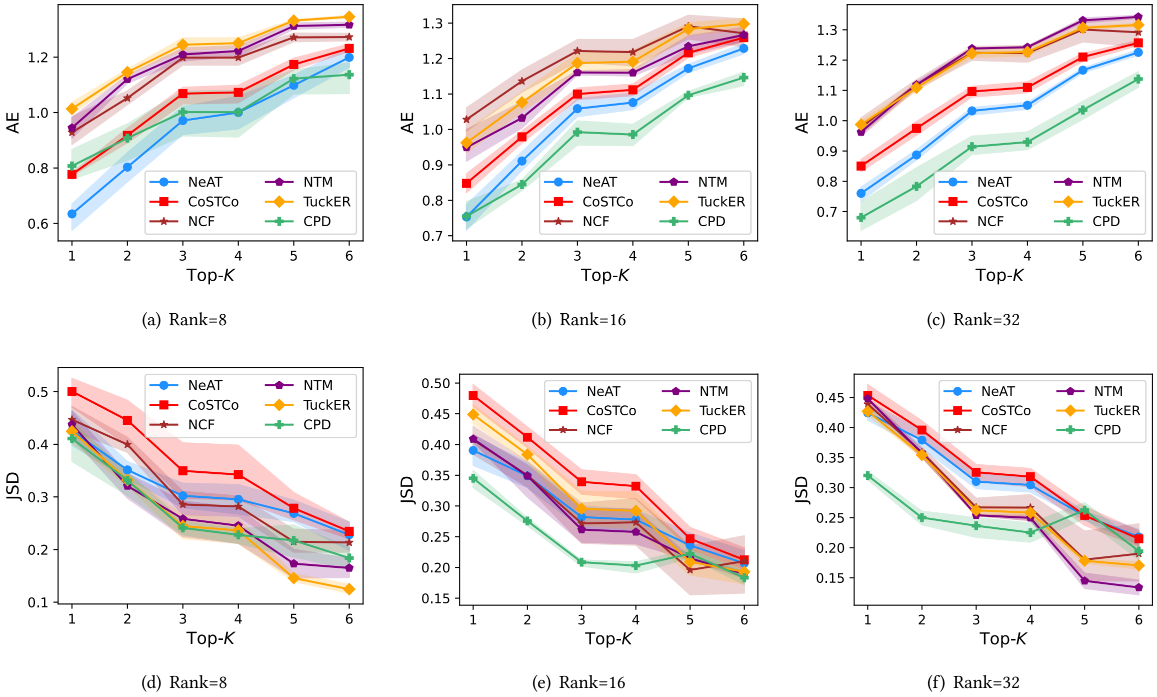
$$\mathbf{p}_r^{(n)} = \frac{1}{K} [\mathbf{p}_r^{(n)}(1) \cdots \mathbf{p}_r^{(n)}(M)], \tag{12}$$

where $\mathbf{p}_r^{(n)}(m) = |\{i_n \in I_n' | l_{i_n} = L_m\}|$ indicates the number of indices having the label L_m and I_n' indicates a set of indices of top-k values in $\mathbf{a}_r^{(n)}$. To measure the distance between two top-K label distributions across the different modes, we use Jenson-Shannon divergence (JSD), which is defined as follows. For n and n' ($n \neq n'$),

$$JSD(\mathbf{p}_r^{(n)} || \mathbf{p}_r^{(n')}) = \frac{1}{2} D(\mathbf{p}_r^{(n)} || Q) + \frac{1}{2} D(\mathbf{p}_r^{(n')} || Q), \tag{13}$$

where $Q = \frac{1}{2}(\mathbf{p}_r^{(n)} + \mathbf{p}_r^{(n')})$.

We varies K according to the size of each mode and use six settings of top-K such as [5, 10, 3], [15, 30, 3], [30, 50, 5], [30, 100, 5], [50, 100, 10], and [100, 1000, 10]. Figure 7 demonstrates the comparison of NeAT with the baselines in terms of Average Entropy (AE) and JSD distance of top-K label distribution over ranks 8 to 32. NeAT shows the second-best AE and a lower JSD compared to CoSTCo, which shows the best AE among neural tensor baselines; a lower value of AE indicates that for NeAT, just like CPD, each rth component is well separated with respect to true labels and a lower value of JSD indicates that rth components across the modes learns similar distribution. For baselines such as NTM, NCF and TuckER, a higher AE and lower JSD indicates that distributions learned by respective models in each component are not separable. As signaled by higher AE, but the lower JSD indicated that differences between label distributions across rth components are small. This implies that even though all models learn similar clusters over rth components , i.e. co-clusters, those co-clusters aren't easily identifiable based on the labels for high entropy methods like NTM, NCF and TuckER.



(a) Rank=8

(b) Rank=16

(c) Rank=32

(d) Rank=8

(e) Rank=16

(f) Rank=32

Figure 7: Comparison of NeAT and baselines in terms of AE and JSD of top-K label distribution. The lower, the better. NeAT shows the second-best lower AE among baselines across the different K and lower JSD between rth component across the mode. This indicates that the components extracted in NeAT are easier to interpret since they are easier to separate from each other. Note that each xtick indicates a mode-wise top-K settings and K values increase toward the right of the x-axis.