	UserBehavior
# of users	969,529
# of items	4,162,024
# of categories	9,439
# of interactions	100,020,395
# of samples	**tens of billions**

Figure 1: Dataset size summary

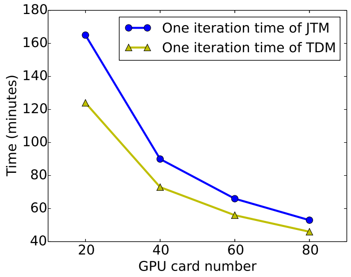

Figure 2: Time cost

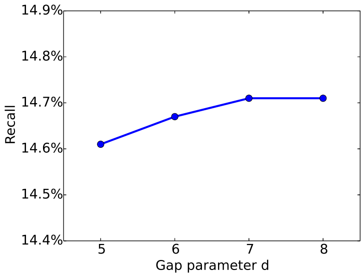

Figure 3: Parameter sensitivity

Thank all the reviewers for your affirmation and insightful comments. Here is the response to some of your concerns.

To Reviewer #1 & #2

As that the time cost is a common concern, we do further experiments to quantify the time cost (detailed time complexity analysis is given in Line 121 of the supplementary material). Here we evaluate the running time with UserBehavior dataset, the size of which is given in Figure 1. The time cost of one iteration (including one epoch model training and once tree learning) time w.r.t. the GPU card number are given in Figure 2. We can observe that the JTM's running time decreases rapidly with the increase of GPU usage, which indicates that JTM is scalable enough for industrial applications, though involves iterative training. Especially, the time cost of JTM approaches to TDM with the increase of card number, while achieving significantly better performance. **JTM has already been fully deployed in the production system. The model and tree structure are daily updated, serving billions of impressions every day.**

To Reviewer #1

Q: It would be helpful to have an intuitive explanation between the proposed JTM and the existing TDM. It would be helpful to have a direct comparison between the tree building steps between JTM and TDM.

JTM addresses the key problem in large-scale recommendation, i.e., how to optimize user representation, user preference prediction and tree structure under a global objective. JTM proposes a **unified framework** to integrate the optimization of these three key factors, while they are optimized separately in TDM. The contribution and improvements are affirmed by the experimental results. JTM optimizes the tree structure by solving the combinatorial optimization problem $\max_{\pi} -\mathcal{L}(\theta, \pi)$ under the unified framework, while TDM uses an intuitive clustering to learn the tree.

Q: It would be good to talk about some of the more practical aspects and the algorithm's parameter sensitivity.

Some of the detailed practical aspects are given in the supplementary material considering the space limit. For example, we use complete binary tree in the experiments, the height of which only depends on the size of item corpus. As for the algorithm's parameter sensitivity, we do additional experiments to evaluate the sensitivity w.r.t. different tree learning gap d (used in Algorithm 2) and the results are in Figure 3. The results indicate that JTM is fairly stable w.r.t. d.

Q: It seems that the clustering algorithm outperforms JTM in the first few iterations, so would be curious about the intuitive explanation why thats the case.

In JTM, the proposed approximate tree learning algorithm involves a *lazy strategy*, i.e., try to reduce the degree of tree structure change in each iteration (details are in Line 39-40 of the supplementary material). That's why the recall metric of JTM doesn't increase as quickly as clustering in the first two iterations. However, the recall metric of JTM keeps increasing and converges to much better final results compared to clustering.

To Reviewer #3

Q: It would've been useful to compare against variants of TDM other than the basic DNN version. It would be interesting to understand how the gap parameter d affects the performance.

We use the basic DNN version of TDM (in Figure 1 of the supplementary material) in all offline comparison, since JTM's merits do not rely on specific network structure. To verify this, we do further experiments with TDM's attention-DNN variant. The recall results of TDM and JTM in UserBehavior dataset (introduced in Figure 1) are **13.07%** and **18.85%** respectively, which is an even more significant improvement. As for the gap parameter d, we evaluate several values other than 7. The results in Figure 3 indicates that the performance is not sensitive w.r.t. the choice of d.