We would like to express our sincere gratitude to the reviewers for providing their valuable feedback. We are able to collectively address only major comments below, but we will thoroughly implement all the comments in a revision.

[R1,R2,R3,R4]-1 (*Generalization to C clusters and G groups*): In fact, we could derive the minimal sample complexity for the generalized setting, although not included in the current draft for illustrative purpose in light of space limitation:

$$mnp^\star = \frac{1}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\max\left\{\frac{GC}{G-R+1}m\log m, \frac{n\log n-\frac{n^2}{GC}I_g}{\delta_g}, \frac{n\log n-\frac{n^2}{GC}I_{c1}-\frac{(G-1)n^2}{GC}I_{c2}}{\delta_c}\right\}, \quad (1)$$

where the set of G rating vectors in each cluster are spanned by any subset of R vectors in the same set. Note that for $(C,G,R)=(2,3,2)$, the bound in (1) reduces to the result of Thm. 1. This generalization will be added to the revision.

[R1,R2,R3,R4]-2 (*Experiments are conducted on real graphs yet on synthetic ratings*): We used this real-synthetic mixed dataset only for the purpose of corroborating our theory at least under real-graph settings, as in [39, 41]. However, it can also be evaluated on purely real data settings (as suggested by R4) by slightly modifying some components in our 4-phase algorithm. For instance, we could actually make a slight change intended for a more realistic setting in which ratings are real and noise is Gaussian (see **[R1]** below for details), and found this modified algorithm working well for the realistic setting. We will clarify this point together with further experiments on purely real datasets in a revision.

[R1,R2,R3] (*Improvement over [39, 40] offered by exploiting the hierarchical graph structure; Part 1*): Remark 1 focuses on the perfect clustering/grouping regime in which user affiliations are successfully revealed. Even in this regime, we still need to estimate four rating vectors $(v_1^A, v_2^A, v_1^B, v_2^B)$. Hence, as R3 assumed, the problem boils down to four separate subproblems *if the hierarchical structure is ignored*. Notice that under *random sampling* of our assumption, the recovery of each vector of length m requires $m\log m$ observations due to the coupon-collecting effect, yielding to $4m\log m$ samples. This can readily be obtained by [39, 40] which do not exploit the *hierarchical structure*. One key observation here is that some measurements associated with (v_3^A, v_3^B) are completely ignored although they can serve to decode $(v_1^A, v_2^A, v_1^B, v_2^B)$. For example, one can decode (v_1^A, v_2^A, v_3^A) *only with any two* of the three vectors due to the linear dependency $v_3^A = v_1^A \oplus v_2^A$, which forms the basis of the *hierarchical structure*. This is exactly what our information-theoretic results and a corresponding efficient algorithm exploit. We found this exploitation is translated to $\frac{4}{3}$ improvement, thus yielding $3m\log m$ sample complexity. We will provide this discussion in a revision.

[R1,R2] (*Improvement over [39, 40] offered by exploiting the hierarchical graph structure; Part 2*): Remark 3 focuses on the limited-clustering regime in which the hierarchical graph information is scarce. The corresponding optimal sample complexity, that reads as $(n\log n-\frac{1}{6}n^2I_{c1}-\frac{1}{3}n^2I_{c2})/\delta_c$, cannot be retrieved from [39,40] since hierarchical graph structure is not exploited in these works. We will provide further details in a revision.

[R2,R4] (*Motivation of hierarchical clustering in recommender systems*): In real-world recommender systems, both item preferences and user preferences are shown to exhibit hierarchical structures [1]. For instance, users within the same cluster can be further divided into sub-clusters (groups) with similar ratings. We will mention this in a revision.

[R1] (*Intuition behind the XOR dependency among rating vectors*): As you may imagine, we adopt this simplified finite-field model only for the purpose of making an initial step towards a more generalized and realistic model. Fortunately, characterizing the optimal sample complexity under the simple model could also shed insights into developing a *universal and model-free* algorithm that is pertinent to any problem setting as long as some slight modification is made. In order to demonstrate this, we now considered a practical scenario in which ratings are real (for which linear dependency between rating vectors is well-accepted) and observation noise is Gaussian. In this setting, the *detection* problem (under the current model) will be replaced by an *estimation* problem. So we update Alg. 1 to incorporate an MLE of the rating vectors; and modify the local refinement criterion on Line 8 in Alg. 2 to find the group that minimizes some properly-defined distance metric between the observed and estimated ratings. We also conducted an experiment (similar to Fig. 2-d) on semi-real data under real field and Gaussian noise, and we found our algorithm still achieves superior performance over the state of the arts. We will include all of these in a revision.

[R2] (*Re. the worst-case error probability*): $\mathcal{M}^{(\delta)}$ is the set of ground-truth matrices M subject to $\delta := \{\delta_g, \delta_c\}$. Hence, there exists indeed ground-truth for the target matrix. Since the error probability may vary depending on different choices of M (some matrices may be harder to estimate), we employ a conventional minimax approach wherein the goal is to minimize the maximum error probability. For clarification, we will elaborate on this in a revision.

[R3]-1 (*Poly-time complexity of our algorithm and its runtime*): In fact, we demonstrated in supplementary (Sec. 4.3, Page 19) that our algorithm runs faster over state-of-the-arts, with few exceptions. We will move this important part to the main body in a revision. Please note that the complexity bottleneck is in Phase 1 (exact clustering), as it relies upon [54], exhibiting $\mathsf{poly}(n)$ runtime. Recently, we have improved our algorithm so as to work optimally even under almost exact (i.e. weak) clustering, yielding $O(|E|\log n)$ runtime [76]. In return, we modified Phase 4 so that the local iterative refinement is applied on cluster affiliation, as well as group affiliation and rating vectors. Hence, the improved overall runtime now reads $O((|\Omega|+|E|)\log n)$. The details of the improved algorithm will be provided in a revision.

[R3]-2 (*Missing references*): Thanks for pointing out the two related papers. We will cite them in a revision.

[R3]-3 (*Mapping between Y and Z*): The adopted mapping (0 to +1, +1 to -1, and * to 0) is one standard way. One can also use another mapping, as you suggested. Yes, we can start with Z instead of Y. We will do so in a revision.

[1] S. Wang, J. Tang, Y. Wang, and H. Liu, "Exploring implicit hierarchical structures for recommender systems," IJCAI, 2015.