We thank the reviewers for the positive recommendation for the paper, and also for the constructive feedback. Since, no critical concern has been raised, we list our answers for the different comments/suggestions made by the reviewers.

Reviewer 1.

More information on the datasets used in the experiment section. The original MNIST dataset is a classification dataset split into two parts for training and testing. For the purpose of clustering, we propose to use the test set. This dataset contains 10000 examples that are roughly equally distributed between the ten digits. In our first experiment (MNIST 1vs7), we select all the 1 and 7 (2163 examples in total). In the second experiment (MNIST 10), we subsample, without replacement, 2000 examples from the whole test set. In both case, we use the Gaussian similarity (See Section F.3 in the supplementary) to generate the comparisons and we randomly and uniformly draw, without replacement, between $n(\ln n)^2$ and $n(\ln n)^4$ comparisons to be observed by the different approaches. The food dataset contains 100 examples and 190376 triplet comparisons. Unfortunately, there is no ground truth and, thus, quantitatively assessing the quality of the obtained partitions is difficult. We will include this in the final version.

Full algorithm in the main text. We will utilise the additional page in camera ready version to add this.

Reviewer 2.

The empirical results for real (car) data are quite weak . . . could potentially be addressed by considering variations on the model. + Robust reconstruction (Moitra et al) The suggestion of considering semirandom models to address mis-specification is very helpful. We will consider (and mention) this as part of future work. Furthermore, we note that the comparisons were obtained from a crowd with view different from the expert who labelled the cars. This would call for modelling noise in crowdsourcing. For example, rather than the type of cars, the crowd might focus on its color. This might explain the differences between our partition based on comparisons and the ground truth. This might also explain why SPUR only detected 2 clusters instead of 3 (as noted by Reviewer 3).

Further details could be provided on for the MNIST data (1v7 and full-10) on how the comparison triplets/quadlets were generated. Please see our answer to the first question of Reviewer 1.

AddS-3/4 kernels as “sufficient statistic" for the model. We thank the reviewer for this suggestion. As noted in the review, this statement is difficult to formalise. We will think about this and, if possible, add it in the follow-up work.

Reviewer 3.

a) Paper uses the lower bound notation (i.e., the Ω function) when presenting upper bounds. We thank the reviewer for this comment and we agree that some of our notations were not correct. We will check all of them and make the appropriate changes.

b) Meaning of lines 68-69 not clear. In order to accurately select which partition an item x belongs to, we need to have some information about it. Otherwise, x can only be assigned to a random cluster. Since, in our setting, we only have access to the examples through comparisons, we need to observe at least one comparison for each example. Hence, $\Omega(n)$ comparisons is a necessary requirement to cluster n items. We will clarify this in the final version.

c) In general, your algorithm can use both triples and quadruples. Yes, but this usually does not happen as in practice a single type of comparison is obtained from the crowd (typically triplets). In fact, we are not aware of any real world scenario where multiple types of comparisons are collected simultaneously.

d) lines 104-106 not clear in context of passive comparisons. These sentences have only a mild underlying assumption, that is, all w_{ij} are distinct (this is not strictly necessary, but it slightly simplifies the analysis). Assuming distinctness, there are no further assumptions. If the triple i, j, r with i as reference *is observed*, then obviously either $w_{ij} > w_{ir}$ or $w_{ij} < w_{ir}$. If the reviewer meant that both could occur when we have multiple observations of the same comparison. We note that this is rare in practice but, in the experiments, we simply follow Ghoshdastidar et al (2019) and use a majority vote.

f) Should the excess value of inter-cluster similarity be $\Omega(p\epsilon\delta\binom{n}{2})$ instead of just $p\epsilon\delta\binom{n}{2}$? The value $p\epsilon\delta\binom{n}{2}$ for quadruplets is exact. In fact, Equation (21) in Appendix C.2 also gives the exact value for the triplet setting as $p\epsilon\delta(n-2)$. We will remove the unnecessary $\Omega(\cdot)$.

g) Remark 1 also seems to need the assumption regarding similar cluster sizes. Yes, all the remarks are stated assuming similar cluster sizes. We quickly mention this in the last paragraph on page 5. We will make this more explicit.

h) it would be better to indicate whether k_λ should be defined as the ceiling or floor function of trace(X_λ). k_λ is the rounded function of trace(X_λ), that is the closest integer.

Other textual comments. Thank you, we will address them in the camera ready version.