We thank all the reviewers for their time and for their thoughtful comments. We agree with all that was said and will do our best to address it in the final version. In particular, we will address the valuable suggestions on the presentation (including adding pseudocode) and we agree that it would be interesting to run our algorithms on larger datasets, which will likely increase the gap in running time between the classic n^2 implementation and our algorithm, as the theory predicts and our first experiments show. We will run the experiments on KDDCUP99 and NEWSGROUP and report them in the final version. The reason we did not prioritize this initially is that, to us, the main value of the experiments is a proof of concept that our notion of approximation leads to good clusters (to be discussed more below) rather than to highlight the speedup (which directly follows from the theoretical gap in complexity and the efficiency of LSH in practice).

The focus of this response will be to discuss the following important concern raised by reviewer #2. We will discuss a few points that were mentioned too briefly in the paper (or not at all), but that will be included in the full version.

On a theoretical justification for our notions of approximation (a concern raised by reviewer #2)

The approximate Average-Linkage notion that we define (γ-AL) guarantees that at every step, the merged pair is γ-close to the best one. But can we prove any guarantees on the quality of the final tree? Will it be "close" to the output of (exact) AL? (Same is true for Ward's, but let us focus on AL in this response.)

One approach that we have considered that also seems to be what the reviewer has in mind is to look at certain objective functions that measure the quality of a hierarchical clustering tree, and compare the guarantees of AL and our γ-AL w.r.t. these objective functions. Such functions were proposed by [2], and [4] (and by [3] for similarity graphs). It is likely that one can prove that γ-AL is guaranteed to give a solution that is no worse than an $O(\gamma)$ factor from the guarantees of (exact) AL w.r.t. to these objective functions. However, such a theorem may not have much value because (as shown by Charikar et al. [1]) the guarantees of AL are no better than those of a random recursive partitioning of the dataset. Therefore, such a theorem will only prove that γ-AL is not-much-worse than random, which dramatically understates the quality of γ-AL. In fact, in our experiments with a standard classification task, γ-AL is very close to AL and is *much* better than random (random has a $1/k$ success rate, which is 0.3 or less, while ours achieves $0.5 - 0.8$).

Another approach would be to prove theorems pertaining to an objective function for HC that offers the guarantee that given two trees, if their costs are close then the structures of their HCs are similar. Unfortunately, we are not aware of any such objective functions (this is also the case for flat clusterings such as k-median, k-means, etc.). In particular, with the functions of [2, 4] the trees output by AL and by a random recursive partitioning have the same cost, while their structure may be very different.

Besides the empirical evidence which, despite being on a small dataset, we find to be a promising proof of concept that our approximate notions make sense, let us mention two more reasons (that we will add to the paper) for the value of our algorithms:

First, our algorithm is essentially a reduction to Approximate Nearest Neighbor (ANN) queries, and ANN queries (using LSH for example) perform very well in practice. In fact, on real world inputs, the algorithm often identifies the *exact* nearest neighbor and then performs the same merge as in AL.

Second, we can provide a theoretical analysis of the following form in support of γ-AL. It is known that if the input data is an ultrametric, then AL (and also Single-Linkage or Complete-Linkage) does recover the underlying ultrametric tree (see e.g.: Cohen-Addad et al.) . Now, assume that the ultrametric is 'clear' in the sense that if $d(a,b) > d(a,c)$ then $d(a,b) > \gamma d(a,c)$ for some constant γ. In this case, our algorithm will provably recover the ultrametric in $n^{1+O(1/\gamma)}$ time, whereas AL would need $\Omega(n^2)$ time. Notably, in this setting, obtaining an $O(1)$-approximation w.r.t. the objective functions of [2, 4] does not mean that the solution is close to the ultrametric tree.

We wish to sincerely thank the reviewers and the PC again for their time and help in improving the quality of this work.

References

[1] M. Charikar, V. Chatziafratis, and R. Niazadeh. Hierarchical clustering better than average-linkage. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2291–2304. SIAM, 2019.

[2] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu. Hierarchical clustering: Objective functions and algorithms. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 378–397. SIAM, 2018.

[3] S. Dasgupta. A cost function for similarity-based hierarchical clustering. *arXiv preprint arXiv:1510.05043*, 2015.

[4] B. Moseley and J. Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In *Advances in Neural Information Processing Systems*, pages 3094–3103, 2017.