A. Appendix

A.1. Analysis of Loss Function for ExpLink

Fact 1 *Given an injective linkage function Ψ^{α}, performing greedy* HAC *with* ExpLink-α *and dissimilarity function f_{θ} results in cluster tree with perfect dendrogram purity if the loss, $J(\theta, \alpha)$, given by equation 4 is zero*

Proof. Let $\mathbf{X} = \{x_i\}_{i=1}^{m}$ with ground-truth clusters $\mathcal{C}^{\star} = \{C_i^{\star}\}_{i=1}^{K}$. Let$\mathcal{T}$ be tree build by HAC on $\mathbf{X}$ using linkage function Ψ^{α} and dissimilarity function f_{θ}.

To prove $J(\theta, \alpha) = 0 \implies \texttt{DP}(\mathcal{T}) = 1$, we will prove the contrapositive i.e. $\texttt{DP}(\mathcal{T}) < 1 \implies J(\theta, \alpha) > 0$.

$\texttt{DP}(\mathcal{T})$ is purity of the lowest common ancestor of a pair of points in $\mathcal{T}$, averaged over every pair of points in the same ground-truth cluster. If $\texttt{DP}(\mathcal{T}) < 1$, then $\exists C^{\star} \in \mathcal{C}^{\star}, x_a, x_b \in C^{\star}$ s.t. $\texttt{purity}(\texttt{LCA}(x_a, x_b), C^{\star}) < 1$ i.e. $\texttt{LCA}(x_a, x_b)$ is the root node of an impure subtree in $\mathcal{T}$.

Let $v_{a,b} = \texttt{LCA}(x_a, x_b)$, and let $C_{v_{a,b}}$ be cluster comprised of points at leaves of tree rooted at $v_{a,b}$. Since $\texttt{purity}(v_{a,b}, C^{\star}) < 1$, at least one of $v_{a,b}$'s subtrees is impure. WLOG, suppose the subtree containing x_a is impure. So, there is a descendant v' of $v_{a,b}$ with children v'_l, v'_r such that $x_a \in C_{v'_l}, C_{v'_l} \subset C^{\star}$, and $C_{v'_r} \not\subset C^{\star}$. This means that v' is the first impure ancestor of x_a.

Let j be the smallest round in which such an impure ancestor of any two points in the same ground-truth cluster is created. Let x_a and x_b be these two points. Before round j, either every cluster is pure cluster (i.e., a subset of a ground-truth cluster), or an impure cluster formed by the union of several ground-truth clusterss. If there exists an impure cluster in round j other than those formed by the union of several ground-truth clusters, then it contradicts j being the smallest round in which an impure ancestor of any two points in the same ground-truth cluster is created.

In round j, let the impure merge occur between a pure cluster C_a and a cluster C_b where $\exists C^{\star} \in \mathcal{C}^{\star}, C_a \subset C^{\star}, C_b \not\subset C^{\star}$. Since C_a is a strict subset of $C^{\star}$, there exists at least one more cluster in round j which a strict subset of $C^{\star}$, and hence there exists at least one pure merger in round j. Let C_{a_+,b_+} be the best pure merger available in round j.

Since HAC chooses to merge $C_{a,b}$ in round j over C_{a_+,b_+},

$$\Psi^{\alpha}(C_{a,b}) \leq \Psi^{\alpha}(C_{a_+,b_+})$$

Further, since Ψ^{α} is injective, we have a strict inequality

$$\begin{aligned} &\Psi^{\alpha}(C_{a,b}) < \Psi^{\alpha}(C_{a_+,b_+}) \\ \implies &\Psi^{\alpha}(C_{a_+,b_+}) - \Psi^{\alpha}(C_{a,b}) > 0 \end{aligned}$$

Thus, $J(\theta, \alpha) \geq \max\{0, \Psi^{\alpha}(C_{a_+,b_+}) - \Psi^{\alpha}(C_{a,b})\} > 0$

Loss incurred in round j is greater zero because the pure merger available in round j is worse than best impure merger available in round j.

□

A.2. Comparison to other inference methods

Top-down tree construction methods have been shown to be effective at optimizing unsupervised hierarchical clustering objectives (Dasgupta, 2016). While there is no natural extension of our training objective for these inference methods, we provide an empirical comparison between HAC inference and the recursive sparsest cut (RSC) approach with the dissimilarity function trained using different training procedures. We implement RSC using scikit-learn's spectral clustering (Pedregosa et al., 2011).

Figure 4 shows mean dendrogram purity results for 50 train/test/dev splits. Each row corresponds to a training procedure for learning the dissimilarity function. The HAC column contains the best dendrogram purity for hierarchical clustering using SL, Avg, Comp or Exp linkage, and the RSC column contains dendrogram purity for top-down hierarchical clustering obtained using recursive sparsest cut. The results of this experiments show that the approaches that use an inference procedure aligned with the training procedure (namely the HAC-based approach presented in this paper) are always more performant than RSC.

Obj	Rexa		AMINER	
	HAC	RSC	HAC	RSC
Bst	87.8	74.3	93.6	88.8
MST	88.4	74.8	93.2	88.1
Exp-	88.6	73.1	85.3	79.3
AP	84.6	75.0	93.4	87.9
Trp	89.1	77.2	93.2	87.3
Exp0	89.5	76.6	94.1	81.6
Exp+	88.1	76.3	92.7	81.5
Expα	89.1	75.1	94.1	81.5

Obj	NP Coref		Faces	
	HAC	RSC	HAC	RSC
Bst	60.5	32.9	93.7	69.6
MST	59.1	37.6	95.4	74.7
Exp-	64.3	49.3	94.6	73.6
AP	58.7	39.7	91.3	81.0
Trp	62.2	54.1	91.0	81.0
Exp0	63.5	50.5	91.0	78.5
Exp+	62.8	52.6	90.4	78.7
Expα	63.4	50.4	94.5	72.9

Figure 4. Dendrogram purity results for RSC and HAC with best linkage function for eight training methods.