HIERARCHICAL CLUSTERING USING RANDOMLY SELECTED MEASUREMENTS

Brian Eriksson

Department of Computer Science, Boston University

ABSTRACT

The problem of hierarchical clustering items from pairwise similarities is found across various scientific disciplines, from biology to networking. Often, applications of clustering techniques are limited by the cost of obtaining similarities between pairs of items. While prior work has been developed to reconstruct clustering using a significantly reduced set of pairwise similarities via adaptive measurements, these techniques are only applicable when choice of similarities are available to the user. In this paper, we examine reconstructing hierarchical clustering under similarity observations at-random. We derive precise bounds which show that a significant fraction of the hierarchical clustering can be recovered using fewer than all the pairwise similarities.

1. INTRODUCTION

Hierarchical clustering based on pairwise similarities arises routinely in a wide variety of engineering and scientific problems. These problems include inferring gene behavior from microarray data [1], Internet topology discovery [2], and advertising [3]. Often times there is a significant cost associated with obtaining each similarity value. This cost can range from computation time to calculate each pairwise similarity (*e.g.,* phylogenetic tree reconstruction using amino acid sequences [4]) to measurement load on the system under consideration (*e.g.,* Internet topology discovery using tomographic probes [2]). In addition, situations where the similarities require an expert human to perform the comparisons results in a significant cost in terms of time and patience of the user (*e.g.,* human perception experiments in [3]).

Prior work has attempted to develop efficient hierarchical clustering methods [5, 6], but many proposed techniques are heuristical in nature and do not provide any theoretical guarantees. Derived bounds are available in [7] which robustly finds clusters down to $O(\log N)$. The main limitation to this approach is that it is only applicable for problems where the user has control over the specific pairs of items to query and when the pairwise similarities are acquired in an online fashion (*i.e.,* one at a time, using past information to inform future samples). In many situations, either this control is not available to the user, or a subset of similarities are acquired in a batch setting (*e.g.,* recommender systems problems [8]). This motivates resolving the hierarchical clustering given a selected number of pairwise similarities observed at-random, where adaptive control is not available. Specifically, we look to answer the following question: *How many similarities observed at-random are required to reconstruct the hierarchical clustering?* The work in [9] developed a clustering technique to resolve clusters down to size $O(N)$ from random observations, in contrast we look to use off-the-shelf clustering methodologies and reconstruct the clustering hierarchy (to some pruning) exactly with high probability.

While results in [7] indicate that resolving the *entire* hierarchical clustering using a sampling at-random regime requires effectively all the pairwise similarities, we find that a significant fraction of the clustering hierarchy can be resolved accurately. Specifically, we resolve the similarity sampling rate required given a desired level of clustering resolution. The only restriction we will place on the observed pairwise similarities are that they satisfy the *Tight Clustering* (TC) condition, which states that intracluster similarity values are greater than intercluster similarity values. This condition is required for any minimum-linkage clustering procedure, and can commonly be found underlying branching processes where the similarity between items is a monotonic increasing function of the distance from the branching root to their nearest common branch point (such as clustering resources in the Internet [10]), or when similarity is defined by density-based distance metrics [11]. Our results show that to find clusters of size $O\left(N^{\beta}\right)$, where $\beta \in (0,1)$, significantly fewer than all the pairwise similarities are required.

2. HIERARCHICAL CLUSTERING AND NOTATION

Let $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ be a collection of N items which has an underlying *hierarchical clustering* denoted as $\mathcal{T}$.

Definition 1. *A **cluster** $\mathcal{C}$ is defined as any subset of $\mathbf{X}$. A collection of clusters $\mathcal{T}$ is called a **hierarchical clustering** if $\cup_{\mathcal{C}_i \in \mathcal{T}} \mathcal{C}_i = \mathbf{X}$ and for any $\mathcal{C}_i, \mathcal{C}_j \in \mathcal{T}$, only one of the following is true **(i)** $\mathcal{C}_i \subset \mathcal{C}_j$, **(ii)** $\mathcal{C}_j \subset \mathcal{C}_i$, **(iii)** $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$.*

Without loss of generality, we will consider $\mathcal{T}$ as a complete binary tree, with N leaf nodes and where every $\mathcal{C}_k \in \mathcal{T}$ that is not a leaf of the tree, there exists proper subsets $\mathcal{C}_i$ and $\mathcal{C}_j$ of $\mathcal{C}_k$, such that $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$, and $\mathcal{C}_i \cup \mathcal{C}_j = \mathcal{C}_k$.

Our measurements will be from $\mathbf{S} = \{s_{i,j}\}$ the collection of all pairwise similarities between the items in $\mathbf{X}$, with

$s_{i,j}$ denoting the similarity between x_i and x_j and assuming $s_{i,j} = s_{j,i}$. The similarities must conform to the hierarchy of $\mathcal{T}$ through the following sufficient condition.

Definition 2. *The triple* $(\mathbf{X}, \mathcal{T}, \mathbf{S})$ *satisfies the* ***Tight Clustering (TC) Condition*** *if for every set of three items* $\{x_i, x_j, x_k\}$ *such that* $x_i, x_j \in \mathcal{C}$ *and* $x_k \notin \mathcal{C}$*, for some* $\mathcal{C} \in \mathcal{T}$*, the pairwise similarities satisfies,* $s_{i,j} > \max(s_{i,k}, s_{j,k})$.

In words, the TC condition implies that the similarity between all pairs within a cluster is greater than the similarity with respect to any item outside the cluster. Under the TC condition, the tree found by agglomerative clustering will match the true clustering hierarchy, $\mathcal{T}$. Minimum-linkage agglomerative clustering [12] is a recursive process that begins with singleton clusters (*i.e.,* the N individual items to be clustered). At each step of the algorithm, the pair of most similar clusters associated with the largest observed pairwise similarity are merged. The process is repeated until all items are merged into a single cluster. The main drawback to this technique is that it requires knowledge of all $O\left(N^2\right)$ pairwise similarities value (*i.e.,* all values must be known to find the maximum), therefore this methodology will be infeasible for problems where N is large, or where there is a significant cost to obtaining each similarity.

To reduce the measurement cost, we consider an incomplete observation of pairwise similarities. For our specific model, we define the indicator matrix of similarity observations, Ω, such that $\Omega_{i,j} = 1$ if the pairwise similarity $s_{i,j}$ has been observed and $\Omega_{i,j} = 0$ if the pairwise similarity $s_{i,j}$ is not observed (*i.e.,* unknown). The pairwise similarities are observed *uniformly at-random*, defining the similarity observation matrix as,

$$P(\Omega_{i,j} = 1) = p \quad : \forall i, j \tag{1}$$

For some probability, $p > 0$.

To reconstruct the clustering from incomplete measurements, this paper focuses on a slightly modified version of minimum-linkage agglomerative clustering. This process can be considered off-the-shelf agglomerative clustering where the pairwise similarities not observed are simply ignored (*i.e.,* unobserved similarities are zero-filled).

2.1. Canonical Subproblem

The intuition for why we can use an incomplete subset of pairwise similarities can be seen with the canonical subproblem where a single cluster $\mathcal{C}$ is split into two subcluster $\mathcal{C}_L, \mathcal{C}_R$ (where $\mathcal{C} = \mathcal{C}_L \bigcup \mathcal{C}_R$ and $\mathcal{C}_L \bigcap \mathcal{C}_R = \emptyset$). In order to properly resolve the two clusters, it is necessary for the agglomerative clustering algorithm to have enough pairwise similarities to make an informed decision as to which items to cluster together. Using this notation, we define a sampling graph, $\mathcal{G}$, resolved from the pairwise similarity observation matrix Ω.

Definition 3. *Consider the* $N \times N$ *pairwise similarity observation matrix* Ω*, then the* ***sampling graph*** $\mathcal{G} = \{\mathbf{V}, \mathbf{E}\}$ *is defined as graph with* $|\mathbf{V}| = N$ *nodes where the edge* $e_{i,j} = 1$ *if* $\Omega_{i,j} = 1$ (i.e., *the pairwise similarities was observed) and* $e_{i,j} = 0$ *otherwise (*i.e., *the pairwise similarity was not observed).*

In the context of the sampling graph, $\mathcal{G}$, we can state the following proposition with respect to resolving the canonical subproblem.

Proposition 1. *Consider a cluster* $\mathcal{C}$ *consisting of two subclusters,* $\mathcal{C}_L$ *and* $\mathcal{C}_R$ *(such that* $\mathcal{C}_L \bigcup \mathcal{C}_R = \mathcal{C}$ *and* $\mathcal{C}_L \bigcap \mathcal{C}_R = \emptyset$*). Then, an agglomerative clustering algorithm will resolve the two subclusters if and only if the sampling subgraphs associated with each subcluster (*i.e., $\mathcal{G}_L$ *for cluster* $\mathcal{C}_L$ *and* $\mathcal{G}_R$ *for cluster* $\mathcal{C}_R$*) are both connected.*

The intuition behind this proposition is as follows. Any clustering procedure requires enough information to associate each item with the cluster it belongs to. The minimum-linkage agglomerative clustering requires that each item observe at least a single similarity with another item in that cluster. But this is not enough, as we also require that one of these two items have an observed similarity with another item in the remainder of the cluster (*i.e.,* one of the items in the cluster, not including the two items paired together), and so on until the all the items can be clustered. In terms of the sampling graph, this is the requirement that a path can be found between items in the same cluster (as all of these pairwise similarities will be greater than any other item outside of this cluster, as stated using the TC condition). It is then obvious that a cluster of items will only be returned if the sampling graph is connected between those items. An example of this clustering can be found in Figure 1-(A,B,C).

Alternatively, if a cluster of items is disconnected, into two sampling graph connected components $\mathcal{C}_A, \mathcal{C}_B$ (where $\mathcal{C}_A \bigcup \mathcal{C}_B = \mathcal{C}$ and $\mathcal{C}_A \bigcap \mathcal{C}_B = \emptyset$), then the clustering procedure will not have enough information to merge the items into a single cluster. An example of this incorrect clustering can be found in Figure 1-(D,E,F).

3. MAIN RESULTS

When pairwise similarities are observed uniformly at-random, such that each pairwise similarity is observed with probability p, the resulting sampling graph can be considered a bernoulli random graph (where each edge exists with probability p). Using Proposition 1 and prior work on random graph theory [13], we can state the following theorem.

Theorem 3.1. *Consider the quadruple* $(\mathbf{X}, \mathcal{T}, \mathbf{S}, \Omega(p))$*, where the Tight Clustering (TC) condition is satisfied, and* $\mathcal{T}$ *is a complete (possible unbalanced) binary tree that is unknown. Then, the agglomerative clustering algorithm recovers all clusters of size* $\geq n$ *of* $\mathcal{T}$ *with probability* $\geq (1 - \alpha)$

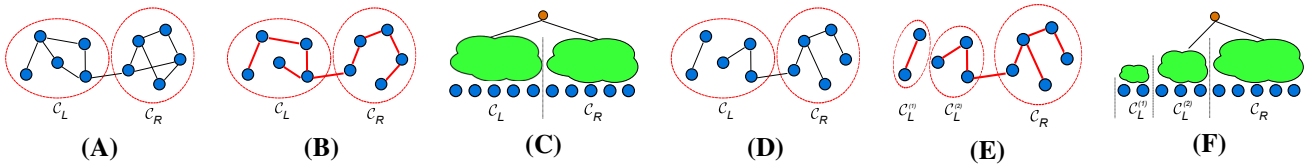

Fig. 1. (A) - Dense sampling graph on two clusters $\mathcal{C}_L, \mathcal{C}_R$, (B) - Example connected path found through the sampling graph using agglomerative clustering, (C) - Resulting correct hierarchical clustering, (D) - Sparse sampling graph on two clusters $\mathcal{C}_L, \mathcal{C}_R$, (E) - Example disconnected path found through the sampling graph using agglomerative clustering, (F) - Resulting incorrect hierarchical clustering.

for $\alpha > 0$ given sampling $\Omega(p)$ satisifies,

$$p \geq \max\left\{1-\left(\frac{\alpha n}{4N}\right)^{2/n}, 1-\left(\frac{\alpha n}{4N\left(1+\frac{\alpha n}{4N}\right)^{n-1}}\right)^{2/n}, 1-\left(\frac{\alpha}{2(N-n)}\right)^{1/n}\right\} \quad (2)$$

Proof. The proof of this theorem follows from the results of Propositions 2 and 3. □

The first component of Theorem 3.1 requires that the sampling probability is large enough that a path can be found with high probability for any collection of items of size $\geq n$.

Proposition 2. *Given a set of N items, then the agglomerative clustering algorithm will recover all $\leq \frac{N}{n}$ leaf clusters of size $\geq n$ of $\mathcal{T}$ with probability $\geq \left(1-\frac{\alpha}{2}\right)$ given the sampling probability satisfies,*

$$p \geq 1 - \min\left\{\left(\frac{\alpha n}{4N}\right)^{2/n}, \left(\frac{\alpha n}{4N\left(1+\frac{\alpha n}{4N}\right)^{n-1}}\right)^{2/n}\right\} \quad (3)$$

Proof. We begin by determining the sampling probability, p, necessary to ensure that a single collection of n items can be clustered. This is equivalent to a bernoulli random graph ($\mathcal{G}_{n,p}$) of size n and probability p being connected. From [13], we bound this probability as,

$$P(\mathcal{G}_{n,p} \text{ is connected}) \geq 1 - q^{n/2}\left(\left(1+q^{\frac{n-2}{2}}\right)^{n-1}-1\right) - q^{n-1}\left(\left(1+q^{\frac{n-2}{2}}\right)^{n-1} - q^{\frac{(n-1)(n-2)}{2}}\right)$$

Where $q = 1-p$. Simplifying in terms of n,

$$P(\mathcal{G}_{n,p} \text{ is connected}) \geq 1 - 2q^{n/2}\left(1+q^{\frac{n}{2}}\right)^{n-1}$$

Considering the entire set of N items, there will be at most $\frac{N}{n}$ leaf clusters to resolve (where each cluster has $\geq n$ items). Therefore, we bound the probability that $\mathcal{G}_{n,p}$ is disconnected as, $2q^{n/2}\left(1+q^{\frac{n}{2}}\right)^{n-1} \leq \frac{\alpha n}{2N}$. Where, using the union bound, we state that all $\leq \frac{N}{n}$ clusters of n items will be connected with probability $\geq 1-\frac{\alpha}{2}$.

Bounding $q^{N/2} \leq C$, with $C > 0$ being a dummy variable, then

$$2q^{n/2}\left(1+q^{\frac{n}{2}}\right)^{n-1} \leq 2q^{n/2}(1+C)^{n-1} \leq \frac{\alpha n}{2N}$$

Therefore, we can solve with respect to the probability of observation, p, and dummy variable C that,

$$p \geq \max\left\{1-C^{2/n}, 1-\left(\frac{\alpha n}{4N(1+C)^{n-1}}\right)^{2/n}\right\} \quad (4)$$

We note that for any choice of α, n, N, the function $1-C^{2/N}$ is monotonically decreasing in C, while $1-\left(\frac{\alpha n}{4N(1+C)^{n-1}}\right)^{2/n}$ is monotonically increasing in C. Therefore, to determine the best choice of C, we find the value where these two functions meet,

$$C^{1/(1-n)} = \left(\frac{\alpha n}{4N}\right)^{1/(1-n)}(1+C)$$

Assuming $C \ll 1$, then $C+1 \approx 1$. Therefore, we use $C \approx \frac{\alpha n}{4N}$. Plugging this value of C into Equation 4 gives us the result. □

While Proposition 2 ensures that there are enough pairwise similarities to determine each leaf cluster (down to size n), to resolve the entire tree structure down to clusters of size $\geq n$ we additionally require enough similarities to determine the connectivity between these clusters.

Proposition 3. *Consider the quadruple $(\mathbf{X}, \mathcal{T}, \mathbf{S}, \Omega(p))$, where the Tight Clustering (TC) condition is satisfied, and $\mathcal{T}$ is a complete (possible unbalanced) binary tree that is unknown. Then, given a set of clusters of size $\geq n$, then the clustering structure of $\mathcal{T}$ pruned to cluster size n will be resolved with probability $\geq \left(1-\frac{\alpha}{2}\right)$ given sampling $\Omega(p)$ satisfies,*

$$p \geq 1 - \left(\frac{\alpha}{2(N-n)}\right)^{1/N} \quad (5)$$

Proof: Consider the clustering structure of $\mathcal{T}$ and a single cluster of size n. At most, there will be $N-n$ other clusters

in $\mathcal{T}$ that must be compared against to construct the clustering hierarchy. Given sampling rate p, then at least one item (out of $\geq n$) must observe a pairwise similarities with an item in another cluster. Therefore, to ensure that every cluster satisfies this with probability $\geq \left(1 - \frac{\alpha}{2}\right)$, using the union bound we require the sampling rate to satisfy $(1-p)^n \leq \frac{\alpha}{2(N-n)}$. Rearranging this term finds the result. □

Combining the results of Propositions 2 and 3, we find the sampling probability rate necessary to ensure with high probability (*i.e.,* $\geq 1 - \alpha$) that all the clusters of size $\geq n$ will be resolved, and the clustering hierarchy connectivity between these clusters can be reconstructed. This is shown in Equation 2.

3.1. Measurements Required for Given Cluster Sizes

Using the results from Theorem 3.1, we can state the expected number of pairwise similarity measurements needed to observe clusters down to a specified level. For potentially small clusters of size $O\left(N^\beta\right)$, where $\beta \in (0,1)$, we find the following,

Theorem 3.2. *Consider the quadruple $(\mathbf{X}, \mathcal{T}, \mathbf{S}, \Omega(p))$, where the Tight Clustering (TC) condition is satisfied, and $\mathcal{T}$ is a complete (possibly unbalanced) binary tree that is unknown. Then, the agglomerative clustering algorithm recovers all clusters of size $O\left(N^\beta\right)$ of $\mathcal{T}$ with probability $\geq (1-\alpha)$ given in expectation we observe $O\left(N^\gamma\right)$ similarities at random, where,*

$$\gamma \geq 2 + \frac{1}{\log N} \max \left\{ \log \left(1 - \left(\frac{\alpha}{4}\right)^{\frac{2}{N^\beta}} N^{\frac{2(\beta-1)}{N^\beta}} \right) , \log \left(1 - \left(\frac{\alpha}{2\left(N - N^\beta\right)} \right)^{N^{-\beta}} \right) \right\} \quad (6)$$

The behavior of the term γ can be seen in Figure 2. The full proof of both Theorem 3.2 can be found in [14] and follows from the results in Theorem 3.1.

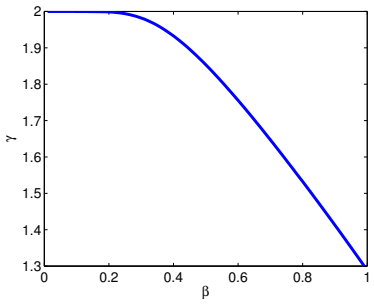

Fig. 2. Derived bound behavior of γ value with respect to β (where with high probability all clusters down to $O\left(N^\beta\right)$ can be resolved given $O\left(N^\gamma\right)$ observed pairwise similarities), for $N = 1000$, $\alpha = 0.05$.

4. REFERENCES

[1] H. Yu and M. Gerstein, "Genomic Analysis of the Hierarchical Structure of Regulatory Networks," in *Proceedings of the National Academy of Sciences*, vol. 103, 2006, pp. 14,724–14,731.

[2] J. Ni, H. Xie, S. Tatikonda, and Y. R. Yang, "Efficient and Dynamic Routing Topology Inference from End-to-End Measurements," in *IEEE/ACM Transactions on Networking*, vol. 18, February 2010, pp. 123–135.

[3] R. K. Srivastava, R. P. Leone, and A. D. Shocker, "Market Structure Analysis: Hierarchical Clustering of Products Based on Substitution-in-Use," in *The Journal of Marketing*, vol. 45, pp. 38–48.

[4] W. Fitch and E. Margoliash, "Construction of Phylogenetic Trees," in *Science*, vol. 155, pp. 279–284.

[5] T. Hofmann and J. M. Buhmann, "Active Data Clustering," in *Proceedings of NIPS*, 1998.

[6] N. Grira, M. Crucianu, and N. Boujemaa, "Active Semi-Supervised Fuzzy Clustering," in *Pattern Recognition*, vol. 41, May 2008, pp. 1851–1861.

[7] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak, "Active Clustering: Robust and Efficient Hierarchical Clustering using Adaptively Selected Similarities," in *Proceedings of AISTATS 2011*, April 2011.

[8] J. Bennet and S. Lanning, "The Netflix Prize," in *KDD Cup and Workshop*, 2007.

[9] M. Balcan and P. Gupta, "Robust Hierarchical Clustering," in *Proceedings of COLT*, July 2010.

[10] R. Ramasubramanian, D. Malkhi, F. Kuhn, M. Balakrishnan, and A. Akella, "On The Treeness of Internet Latency and Bandwidth," in *Proceedings of ACM SIGMETRICS Conference*, Seattle, WA, 2009.

[11] Sajama and A. Orlitsky, "Estimating and Computing Density-Based Distance Metrics," in *Proceedings of ICML*, 2005.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[13] E. Gilbert, "Random Graphs," vol. 30, no. 4. Annuals of Mathematical Statistics, 1959, pp. 1141–1144.

[14] B. Eriksson, "Hierarchical Clustering using Randomly Selected Measurements," in *Boston University Technical Report*, Feburary 2012.