Project Summary

Overview

This proposal presents a distributed algorithm for secure clustering of high dimensional data. The novel algorithm, called Random Projection Hash or *RPHash*, utilizes aspects of locality sensitive hashing (LSH) and multi-probe random projection for computational scalability and linear achievable gains from parallel speed up. The two step approach is data agnostic, minimizes communication overhead, and has a priori predictable computational time. The system is deployable on commercially available cloud resources running the Hadoop (*MR*v2) implementation of MapReduce. The *RPHash* solution will have a wide applicability to a variety of standard clustering applications while this project will focus on a subset of clustering problems in the biological data analysis space. *RPHash* also combats de-anonymization attacks inherently resulting from its algorithmic requirements thus addressing requirements involving the handling and privacy protection of health care data as well as the inherent privacy concerns of using cloud based services. Furthermore, *RPHash* will allow researchers to scale their clustering problems without the need for specialized equipment or computing resources. The proposed cloud processing solution will allow researchers to arbitrarily scale their processing needs using virtually limitless commercial processing resources.

Intellectual Merit

A principle driving force in computational progress results from material and architectural advances in microprocessor design. Many of these advances have been stagnated for linear processing however due to thermal dissipation and energy requirements, resulting in a shift toward parallel multi-processing. Though much has been done to adapt current algorithms for this parallel processing landscape, many solutions tend to be a' posteriori methods favoring empirical speedup over long term scalability. In this proposal, we develop a method for an algorithm central to data analysis, designed expressly for parallel multi-processing systems. In addition, our algorithm addresses often overlooked communication bottlenecks in parallel design, through use of side channel synchronization based on mathematically generative groups and probabilistic approximation. A favorable side effect of the probabilistic approximation results in a possible solution to de-anonymization attacks[**?**][**?**] on user data.

Broader Impacts

Clustering has long been the standard method used for the analysis of labeled and unlabeled data. Clustering's effects intrinsically identify the latent models underlying the distributions of objects in a dataset, often unattainable through standard statistical methods. Single pass, data intensive statistical methods are often the primary workhorses for parallel database processing of business logic and other domains, while clustering is often overlooked due to scalability ocncerns and confusion caused by the wide variety of available distributed algorithms [**?**]. A multitude of surveys [**?**] have been made available comparing different aspects of clustering algorithms, such as accuracy, complexity and application domain. Fields in health and biology are benefited by data clustering scalability. Such fields as Micro Array clustering, Protein-Protein interaction clustering, medical resource decision making, medical image processing, and clustering of epidemiological events all serve to benefit from larger dataset sizes. Furthermore, the proposed method provides data anonymization through destructive manipulation of the data preventing de-anonymization attacks beyond standard best practices database security methods.