Distributed Data Mining for Astronomy Catalogs

Chris Giannella*  Haimoti Dutta†  Kirk Borne‡  Ran Wolff§  Hillol Kargupta¶

Abstract
The design, implementation, and archiving of very large sky surveys is playing an increasingly important role in today’s astronomy research. However, these data archives will necessarily be geographically distributed. To fully exploit the potential of this data, we believe that capabilities ought to be provided allowing users a more communication-efficient alternative to multiple archive data analysis than first down-loading the archives fully to a centralized site.

In this paper, we propose a system, DEMAC, for the distributed mining of massive astronomical catalogs. The system is designed to sit on top of the existing national virtual observatory environment and provide tools for distributed data mining (as web services) without requiring datasets to be fully down-loaded to a centralized server. To illustrate the potential effectiveness of our system, we carry out a case study using distributed principal component analysis (PCA) for detecting fundamental planes of astronomical parameters. In particular, PCA enables dimensionality reduction within a set of correlated physical parameters, such as a reduction of a 3-dimensional data distribution (in astronomer’s observed units) to a planar data distribution (in fundamental physical units). Fundamental physical insights are thereby enabled through efficient access to distributed multi-dimensional data sets.

1 Introduction
The design, implementation, and archiving of very large sky surveys is playing an increasingly important role in today’s astronomy research. Many projects today (e.g. GALEX All-Sky Survey), and many more projects in the near future (e.g. WISE All-Sky Survey) are destined to produce enormous catalogs (tables) of astronomical sources (tuples). These catalogs will necessarily be geographically distributed. It is this virtual collection of gigabyte, terabyte, and (eventually) petabyte catalogs that will significantly increase science return and enable remarkable new scientific discoveries through the integration and cross-correlation of data across these multiple survey dimensions [19]. Astronomers will be unable to fully tap the riches of this data without a new paradigm for *astro-informatics* that involves distributed database queries and data mining across distributed virtual tables of joined and integrated sky survey catalogs [4, 5].

The development and deployment of a National Virtual Observatory (NVO) [23] is a step toward a solution of this problem. However, processing, mining, and analyzing these distributed and vast data collections are fundamentally challenging tasks since most off-the-shelf data mining systems require the data to be down-loaded to a single location before further analysis. This imposes serious scalability constraints on the data mining system and fundamentally hinders the scientific discovery process. Figure 1 further illustrates this technical problem. The left part depicts the current data flow in the NVO. Through web services, data are selected and down-loaded from multiple sky-surveys.

If distributed data repositories are to be really accessible by a larger community, then technology ought to be developed for supporting distributed data analysis that can reduce, as much as possible, communication requirements among the data servers and the client machines. Communication-efficient distributed data mining (DDM) techniques will allow a large number of users simultaneously to perform advanced data analysis without necessarily down-loading large volumes of data to their respective client machines.

In this paper, we propose a system, DEMAC, for the distributed exploration of massive astronomical catalogs. DEMAC will offer a collection of data mining tools based on various DDM algo-

---

*CSEE Department, University of Maryland, Baltimore County, USA cgianel@cs.umbc.edu
†CSEE Department, University of Maryland, Baltimore County, USA hdutta@cs.umbc.edu
‡School of Computational Sciences, George Mason University, USA kborne@gmu.edu
§Computer Science Department, Technion, Israel ran wolff@gmail.com
¶CSEE Department, University of Maryland Baltimore County, USA, hillol@cs.umbc.edu. Also affiliated with AGNIK LLC, Columbia, MD USA.
Figure 1: Proposed data flow for distributed data mining embedded in the NVO.

(a) Current data flow is restricted because of data ownership and bandwidth considerations.

(b) Distributed data mining algorithms can process large amounts of data using a small amount of communication. The users get the data mining output rather than raw data.

rithms. The system would be built on top of the existing NVO environment and provides tools for data mining (as web services) without requiring datasets to be down-loaded to a centralized server. Consider again Figure 1. Our system requires a relatively simple modification—the addition of a distributed data mining functionality in the sky servers. This allows DDM to be carried out without having to down-load large tables to the users’ desktop or some other remote machine. Instead, the users will only down-load the output of the data mining process (a data mining model); the actual data mining from multiple data servers will be performed using communication-efficient DDM algorithms. The algorithms we develop sacrifice perfect accuracy for communication savings. They offer approximate results at a considerably lower communication cost than that of exact results through centralization. As such, we see DEMAC as serving the role of an exploratory “browser”. Users can quickly get (generally quite accurate) results for their distributed queries at low communication cost. Armed with these results, users can focus in on a specific query or portion of the datasets, and down-load for more intricate analysis.

To illustrate the potential effectiveness of our system, we carry out a case study using distributed principal component analysis (PCA) for detecting fundamental planes of astronomical parameters. Astronomers have previously discovered cases where the observed parameters measured for a particular class of astronomical objects (such as elliptical galaxies) are strongly correlated, as a result of universal astrophysical processes (such as gravity). PCA will find such correlations in the form of principal components. An example of this is the reduction of a 3-dimensional scatter plot of elliptical galaxy parameters to a planar data distribution. The explanation of this plane follows from fundamental astrophysical processes within galaxies, and thus the resulting data distribution is labeled the Fundamental Plane of Elliptical Galaxies. The important physical insights that astronomers have derived from this fundamental plane suggest that similar new physical insights and scientific discoveries may come from new analysis of combinations of other astronomical parameters. Since these multiple parameters are now necessarily distributed across geographically dispersed data archives, it is scientifically valuable to explore distributed PCA on larger astronomical data collections and for greater numbers of astrophysical parameters. The application of communication-efficient distributed PCA and other DDM algorithms will likely enable the discovery of new fundamental planes, and thus produce new scientific insights into our Universe.

2 Related Work

2.1 Analysis of Large Data Collections

There are several instances in the astronomy and space sciences research communities where data mining is being applied to large data collections [21]. Some dedicated data mining projects include F-MASS [11], Class-X [9], the Auton Astrostatistics Project [3]. In essentially none of these cases does
the project involve truly DDM [20]. Through a past NASA-funded project, K. Borne applied some very basic DDM concepts to astronomical data mining [8]. However, the primary accomplishments focused only on centralized co-location of the data sources [6, 7].

One of the first large-scale attempts at grid data mining for astronomy is the U.S. National Science Foundation (NSF) funded GRIST [13] project. The GRIST goals include application of grid computing and web services (service-oriented architectures) to mining large distributed data collections. GRIST is focused on one particular data modality: images. Hence, GRIST aims to deliver mining on the pixel planes within multiple distributed astronomical image collections. The project that we are proposing here is aimed at another data modality: catalogs (tables) of astronomical source attributes. GRIST and other projects also strive for exact results, which usually requires data centralization and co-location, which further requires significant computational and communications resources. DEMAC (our system) will produce approximate results without requiring data centralization (low communication overhead). Users can quickly get (generally quite accurate) results for their distributed queries at low communication cost. Armed with these results, users can focus in on a specific query or portion of the datasets, and download for more intricate analysis.

The U.S. National Virtual Observatory (NVO) [23] is a large scale effort funded by the NSF to develop an information technology infrastructure enabling easy and robust access to distributed astronomical archives. It will provide services for users to search and gather data across multiple archives and some basic statistical analysis and visualization functions. It will also provide a framework for new services to be made available by outside parties. These services can provide, among other things, specialized data analysis capabilities. As such, we envision DEMAC to fit nicely into the NVO as a new service.

2.2 Distributed Data Mining DDM is a relatively new technology that has been enjoying considerable interest in the recent past [16]. DDM algorithms strive to analyze the data in a distributed manner without downloading all of the data to a single site (which is usually necessary for a regular centralized data mining system). DDM algorithm naturally fall into two categories according to whether the data is distributed horizontally (with each site having some of the tuples) or vertically (with each site having some of the attributes for all tuples). In the latter case, it is assumed that the sites have an associated unique id used for matching. In other words, consider a tuple $t$ and assume site $A$ has a part of this tuple, $t_A$, and $B$ has the remaining part $t_B$. Then, the id associated with $t_A$ equals the id associated with $t_B$.¹

The NVO can be seen as a case of vertically distributed data, assuming ids have been generated by a cross-matching service. With this assumption, DDM algorithms for vertically partitioned data can be applied. These include algorithms for principal component analysis (PCA), Bayesian network learning, clustering, and supervised classification (see [16] for references).

3 Data Analysis Problem: Analyzing Distributed Virtual Catalogs

We illustrate the problem with two archives: the Sloan Digital Sky Survey (SDSS) [24] and the 2-Micron All-Sky Survey (2MASS) [1]. Each of these has a simplified catalog containing records for a large number of astronomical point sources, upward of 100 million for SDSS and 470 million for 2MASS. Each record contains sky coordinates $(ra, dec)$ identifying the sources’ position in the celestial sphere as well as many other attributes (460+ for SDSS; 420+ for 2MASS). While each of these catalogs individually provides valuable data for scientific exploration, together their value increases significantly. In particular, efficient analysis of the virtual catalog formed by joining these catalogs would enhance their scientific value significantly. Henceforth, we use “virtual catalog” and “virtual table”, interchangeably.

To form the virtual catalog, records in each catalog must first be matched based on their position in the celestial sphere. Consider record $t$ from SDSS and $s$ from 2MASS with sky coordinates $[ra, dec]$ and $[ra, dec]$. Each record represents a set of observations about an astronomical object e.g. a galaxy. The sky coordinates are used to determine if $t$ and $s$ match, i.e. are close enough that $t$ and $s$ represents the same astronomical object. The issue of how matching is done will be discussed later. For each match $(t, s)$, the result is a record $t owtie s$ in the virtual catalog with all of the attributes of $t$ and $s$. As described earlier, the virtual catalog provides valuable data that neither SDSS or 2MASS alone can provide.

DEMAC addresses the data analysis problem of developing communication-efficient algorithms for analyzing user-defined subsets of virtual catalogs.

¹Each id is unique to the site at which it resides; no two tuples at site $A$ have the same id.
The algorithms allow the user to specify a region \( R \) in the sky and a virtual catalog, then efficiently analyze the subset of tuples from that catalog with sky coordinates in \( R \). Importantly, the algorithms we propose do not require that the base catalogs first be centralized and the virtual catalog explicitly realized. Moreover, the algorithms are not intended to be a substitute for exact, centralization-based methods currently being developed as part of the NVO. Rather, they are intended to complement these methods by providing, quick, communication-efficient approximate results to allow browsing. Such browsing will allow the user to better focus their exact, communication-expensive, queries.

**Example 1.** The all data release of 2MASS contains attribute, “\( K \)-band means surface brightness” (\( Kmnb \)). Data release four of SDSS contains galaxy attributes “redshift” (\( rs \)), “petrosian I band angular effective radius” (\( Iae \)) and “velocity dispersion” (\( v_d \)). To produce a physical variable, consider composite attribute “petrosian I band effective radius” (\( Ier \)) formed by the product of \( Iae \) and \( rs \). Note, since \( Iae \) and \( rs \) are both at the same repository (SDSS), then, from the standpoint of distributed computation, we may assume \( Ier \) is contained in SDSS.

A principal component analysis over a region of sky \( R \) on the virtual table with columns \( \log(Ier) \), \( \log(v_d) \), and \( Kmnb \) is interesting in that it can allow the identification of a “fundamental plane” (the algorithms are used to place all variables on the same scale). Indeed, if the first two principal components capture most of the variance, then these two variables define a fundamental plane. The existence of such things points to interesting astrophysical behaviors. We develop a communication-efficient distributed algorithm for approximating the principal components of a virtual table.

**4 DEMAC - A System for Distributed Exploration of Massive Astronomical Catalogs**

This section describes the high level design of the proposed DEMAC system. DEMAC is designed as an additional web-service which seamlessly integrates into the NVO. It consists of two basic services. The main one is a web-service providing DDM capabilities for vertically distributed sky surveys (WS-DDM). The second one, which is intensively used by WS-DDM, is a web-service providing cross-matching capabilities for vertically distributed sky surveys (WS-CM). Cross-matching of sky surveys is a complex topic which is dealt with, in itself, under other NASA funded projects. Thus, our implementation of this web-service would supply bare minimum capabilities which are required in order to provide distributed data mining capabilities.

To provide a distributed data mining service, DEMAC would rely on other services of the NVO such as the ability to select and download from a sky survey in an SQL-like fashion. Key to our approach is that these services be used not over the web, through the NVO, but rather by local agents which are co-located with the respective sky survey. In this way, the DDM service avoid bandwidth and storage bottlenecks, and overcomes restrictions which are due to data ownership concerns. Agents, in turn, take part in executing efficient distributed data mining algorithms, which are highly communication-efficient. It is the outcome of the data mining algorithm, rather than the selected data table, that is provided to the end-user. With the removal of the network bandwidth bottleneck, the main factor limiting the scalability of the distributed data mining service would be database access. For database access we intend to rely on the SQL-like interface provided to the different sky-surveys to the NVO.

We outline here the architecture we propose for the two web-services we will develop.

**4.1 WS-DDM – DDM for Heterogeneously Distributed Sky-Surveys**

This web-service will allow running a DDM algorithm (one will be discussed later) on a selection of sky-surveys. The user would use existing NVO services to locate sky-surveys and define the portion of the sky to be data mined. The user would then use WS-CM to select a cross-matching scheme for those sky-surveys. This specifies how the tuples are matched across surveys to define the virtual table to be analyzed. Following these two preliminary phases the user would submit the data mining task.

Execution of the data mining task would be scheduled according to resource availability. Specifically, the size of the virtual table selected by the user would dictate scheduling. Having allocated the required resources, the data mining algorithm would be carried on by agents which are co-located with the selected sky-surveys. Those agents will access the sky-survey through the SQL-like interface it exposes to the NVO and will communicate with each other directly, over the Internet. When the algorithm has terminated, results would be provided to the user using a web-interface.
4.2 WS-CM — Cross-Matching for Heterogeneously Distributed Sky-Surveys

Central to the DDM algorithms we develop is that the virtual table can be treated as vertically partitioned (see Section 2 for the definition). To achieve this, match indices are created and co-located with each sky survey. Specifically, for each pair of surveys (tables) \( T \) and \( S \), a distinct pair of match indices must be kept, one at each survey. Each index is a list of pointers; both indices have the same number of entries. The \( i^{th} \) entry in \( T \)'s list points to a tuple \( t_i \) and the \( i^{th} \) entry in \( S \)'s list points to \( s_i \), such that \( t_i \) and \( s_i \) match. Tuples in \( T \) and \( S \) which do not have a match, do not have a corresponding entry in either index. Clearly, algorithms assuming a vertically partitioned virtual table can be implemented on top of these indices.

Creating these indices is not an easy job. Indeed, cross-matching sources is a complex problem for which no single best solution exists. The WS-CM web-service is not intended to address this problem. Instead it will use already existing solutions (e.g., the cross-matching service already provided by the NVO), and it will be designed to allow other solutions to be plugged in easily. Moreover, cross-matching the entirety of two large surveys is a very time-consuming job and would require centralizing (at least) the \( ra, dec \) coordinates of all tuples from both.

Importantly, the indices do not need to be created each time a data mining task is run. Instead, provided sky survey data is static (it generally is), each pair of indices only need be created once. Then any data mining task can use them. In particular the DDM tasks we develop can use them. The net result is the ability to mine virtual tables at low communication cost.

4.3 Definitions and Notation

In the next section we describe a DDM algorithm to be used as part of the WS-DDM web service. It assumes that the participating sites have the appropriate alignment indices. Hence, for simplicity, we describe the algorithms under the assumption that the data in each site is perfectly aligned — the \( i^{th} \) tuple of each site match (sites have exactly the same number of tuples). This assumption can be emulated without problem using the matching indices.

Let \( M \) denote an \( n \times m \) matrix with real-valued entries. This matrix represents a dataset of \( n \) tuples from \( \mathbb{R}^m \). Let \( M_j \) denote the \( j^{th} \) column and \( M_i \) denote the \( i^{th} \) entry of this column. Let \( \mu(M_j) \) denote the sample mean of this column, \( \frac{\sum_{i=1}^{n} M_{j}(i)}{n} \). Let \( \text{Var}(M_j) \) denote the sample variance of this column, \( \Sigma_{i=1}^{n} \left( M_{j}(i)-\mu(M_j) \right)^{2} \). Let \( \text{Cov}(M_j, M_k) \) denote the sample covariance of the \( j^{th} \) and \( k^{th} \) columns, \( \Sigma_{i=1}^{n} \left( M_{j}(i)-\mu(M_j) \right) \left( M_{k}(i)-\mu(M_k) \right) \). Note, \( \text{Var}(M_j) = \text{Cov}(M_j, M_j) \).

Suppose this dataset has been vertically distributed over two sites \( S_A \) and \( S_B \). Since we are assuming that the data at the sites is perfectly aligned, then \( S_A \) has the first \( p \) attributes and \( S_B \) has the last \( q \) attributes (\( p + q = m \)). Let \( A \) denote the \( n \times p \) matrix representing the dataset held by \( S_A \), and \( B \) denote the \( n \times q \) matrix representing the dataset held by \( S_B \). Let \( A : B \) denote the concatenation of the datasets i.e. \( M = A : B \). The \( j^{th} \) column of \( A : B \) is denoted \( [A : B]_j \).

Next we describe a communication-efficient algorithm for PCA on \( M \) vertically distributed over two sites. The algorithm easily extends to more than two sites, but, for simplicity, we only discuss the two site scenario. Later we examine its effectiveness through a case study. We have also developed a distributed algorithm for decision tree induction (supervised classification) [12] and are in the process of developing a distributed algorithm for outlier detection.

Following a standard practice in applied statistics, we pre-process \( M \) by normalizing so that \( \mu(M_j) = 0 \) and \( \text{Var}(M_j) = 1 \). This is achieved by replacing each entry \( M_j(i) \) with \( \frac{M_j(i) - \mu(M_j)}{\sqrt{\text{Var}(M_j)}} \). Since both \( \mu(M_j) \) and \( \text{Var}(M_j) \) can be computed without any communication, then normalizing can be performed without any communication. Henceforth, we assume \( \mu(M_j) = 0 \) and \( \text{Var}(M_j) = 1 \).

Let \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0 \) denote the eigenvalues of \( \text{Cov}(M) \) and \( v_1, v_2, \ldots, v_m \) the associated eigenvectors (pairwise orthonormal). The \( j^{th} \) principal direction of \( M \) is \( v_j \). The \( j^{th} \) principal component is denoted \( z_j \) and equals \( M v_j \) (the projection of \( M \) along the \( j^{th} \) direction).

5 Virtual Catalog Principal Component Analysis

PCA is a well-established data analysis technique used in a large number of disciplines: astronomy, computer science, biology, chemistry, climatology, geology, etc. Quoting [14] page 1: "The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the
variation present in the dataset." Next we provide a very brief overview of PCA, for a more detailed treatment, the reader is referred to [14].

The $i$th principal component, $z_j$, is, by definition, a linear combination of the columns of $M$ – the $k$th column has coefficient $v_j(k)$. The sample variance of $z_j$ equals $\lambda_j$. The principal components are all uncorrelated i.e. have zero pairwise sample covariances. Let $Z_r (1 \leq r \leq m)$ denote the $n \times r$ matrix with columns $z_1, \ldots, z_r$. This is the dataset projected onto the subspace defined by the first $r$ principal directions. If $r = m$, then $Z_m$ is simply a different way of representing exactly the same dataset, because $M$ can be recovered completely as $M = Z_m V^T$ where $T$ denotes matrix transpose.3

However, if $r < m$, then $Z_r$ is a lossy lower dimensional representation of $M$. The amount of loss is typically quantified as, $\sum_{j=r+1}^m \lambda_j$, the "proportion of variance" captured by the lower dimensional representation. If $r$ is chosen so that a large amount of the variance is captured, then, intuitively, $Z_r$ captures many of the important features of $M$. So, subsequent analysis on $Z_r$ can be quite fruitful at revealing structure not easily found by examination of $M$ directly. Our case study employs this idea.

To our knowledge, the problem of vertically distributed PCA computation was first addressed by Kargupta et al. [17] based on sampling and communication of dominant eigenvectors. Later, Kargupta and Puttagunta [15] developed a technique based on random projections. Our method is a slightly revised version of this work. We describe a distributed algorithm for approximating $\text{Cov}(A : B)$. Clearly, PCA can be performed from $\text{Cov}(A : B)$ without any further communication.

Recall that $A : B$ is normalized to have zero column sample mean and unit sample variance. As a result, $\text{Cov}(A : B)^T, [A : B]^k = \sum_{i,j} [A : B]^i [A : B]^j \frac{1}{n^2}$ which is the inner product between $[A : B]^i$ and $[A : B]^k$ divided by $n - 1$. Clearly this inner product can be computed without communication when $[A : B]^i$ and $[A : B]^k$ are at the same site (i.e. $1 \leq j, k \leq p$ or $p + 1 \leq j, k \leq p + q$). It suffices to show how the inner product can be approximated across different sites, in effect, how $A^T B$ can be approximated. The key idea is based on the following fact echoing the observation made in [22] that high-dimensional random vectors are nearly orthogonal. A similar result was proved elsewhere [2],

\begin{fact}
Let $R$ be an $\ell \times n$ matrix each of whose entries is drawn independently from a distribution with finite variance and mean zero. It follows that $E[R^T R] = nI_n$ where $I_n$ is the $n \times n$ identity matrix.
\end{fact}

We will use the Algorithm 1 for computing $A^T B$. The result is obtained at both sites.4

\begin{algorithm}
\caption{Distributed Covariance Matrix Algorithm}
\begin{algorithmic}
\State $S_A$ sends $S_B$ a random number generator seed. [1 message]
\State $S_A$ and $S_B$ generate a $\ell \times n$ random matrix $R$ where $\ell$. Each entry is generated independently and identically from any distribution with zero mean and finite variance.
\State $S_A$ sends $RA$ to $S_B$; $S_B$ sends $RB$ to $S_A$. [4m messages]
\State $S_A$ and $S_B$ compute $D = \frac{(RA)^T (RB)}{n}$.
\end{algorithmic}
\end{algorithm}

From Fact 1, it can be seen that $E[D] = A^T E[R^T R] B = A^T B$. Hence, on expectation, the algorithm is correct. However, its communication cost (bytes) divided by the cost of the centralized-based algorithm, $\frac{\ell + n}{\ell m}$, is small if $\ell \ll n$. Indeed $\ell$ provides a "knob" for tuning the trade-off between communication-efficiency and accuracy. Later we present experiments measuring this trade-off.

6 Case Study: Finding Galactic Fundamental Planes

The identification of certain correlations among parameters has lead to important discoveries in astronomy. For example, the class of elliptical and spiral galaxies (including dwarfs) have been found to occupy a two dimensional space inside a 3-D space of observed parameters, radius, mean surface brightness and velocity dispersion. This 2D plane has been referred to as the Fundamental Plane ([10, 18]).

This section presents a case study involving the detection of a fundamental plane among galaxy parameters distributed across two catalogs: 2MASS and SDSS (the problem was described earlier in Example 1). Our goal is to demonstrate that, using our distributed covariance matrix algorithm to approximate the principal components, we can find a very similar fundamental plane as that obtained by applying a centralized PCA. Note that our goal is not to make a new discovery in astronomy. Indeed, the fundamental plane we uncover in our analysis

\begin{figure}
\centering
\includegraphics[width=\textwidth]{example.png}
\caption{Example}
\end{figure}

\begin{table}
\centering
\begin{tabular}{|c|c|c|}
\hline
Parameter & Value & Parameter & Value \\
\hline
Radius & 0.5 & Velocity Dispersion & 100 \\
\hline
Surface Brightness & 1000 & Mass & 10^11 \\
\hline
\end{tabular}
\caption{Example Table}
\end{table}

\begin{itemize}
\item Since $V$ is a square matrix with orthonormal columns, then basic linear algebra shows that $VV^T$ equals the $m \times m$ identity matrix.
\item In the communication cost calculations, we assume a message requires 4 bytes of transmission.
\end{itemize}
is not new. However, we demonstrate that our distributed PCA algorithm could have found the result. Therefore, we argue that our algorithm could find fundamental planes when applied to previously unexplored data. As such, DEMAC could provide a valuable tool for astronomers wishing to explore many parameter spaces across different catalogs for “tighter” planes.

In our study we are interested in measuring the accuracy of our distributed algorithm\(^5\) with respect to the communication it requires. For this purpose, a distributed environment is not necessary. Thus, for simplicity, we used a single machine and simulated a distributed environment. We prepare our test data as follows.

Using the web interfaces of 2MASS\(^6\) and SDSS\(^7\) and the SDSS object cross id tool, we obtain an aggregate dataset involving attributes from 2MASS and SDSS and lying in the sky region between right ascension (ra) 150 and 200 and declination (dec) 0 and 15. The aggregate dataset has the following attributes from SDSS: Petrosian I band angular effective radius \((r_{\text{aer}})\), redshift \((z)\), and velocity dispersion \((v_d)\);\(^8\) and has the following attribute from 2MASS: K band mean surface brightness \((K_{\text{msb}})\).\(^9\) After removing tuples with missing attributes, we have a 1307 tuple dataset with four attributes. We produce a new attribute, logarithm petrosian I band effective radius \((\log(r_{\text{aer}}))\), as \(\log(r_{\text{aer}})\) and a new attribute, logarithm velocity dispersion \((\log(v_d))\), by applying the logarithm to \(v_d\). We drop all attributes except those to obtain the three attribute dataset, \(\log(r_{\text{aer}})\), \(\log(v_d)\), \(K_{\text{msb}}\). Finally, we normalize each column by subtracting its mean from each entry.

We apply PCA directly to this dataset to obtain the centralization based results. Next we treat this dataset as if it were distributed (assuming cross match indeces have been created as described earlier)\(^10\) This can be thought of as a virtual table with attributes \(\log(r_{\text{aer}})\) and \(\log(v_d)\) located at one site and attribute \(K_{\text{msb}}\) at another. Finally, we apply our distributed covariance matrix algorithm and compute the principal components from the resulting matrix. Note, our dataset is somewhat small and not necessarily indicative of a scenario where DEMAC would be used in practice. However, for the purposes of our study (accuracy with respect to communication) it suffices.

The variances of the centralized PCs (times 100) are 61.4, 32.8, and 5.8. This demonstrates a fundamental plane as most of the data variance is captured by the first two PCs (94.2). Hence, the data tends to lie on the plane defined by these two. Figures 2, 3, and 4 compare the variance of the PCs computed by the centralized algorithm with those computed by the distributed algorithm.\(^11\) The x-axes enumerate communication percentage, i.e. the amount of communication used by the distributed algorithm as a percentage of that required to transmit all the data from both sites to a central site (5228 bytes). We see that the variances from the distributed algorithm are reasonably close to the centralized variances; at ten percent communication, the difference between the centralized and distributed variances are approximately -0.8, -4.8, 5.6. Hence, the distributed algorithm would allow a user to detect the fundamental plane (the first two PCs capture 88.6 percent of the total variance).

Recall that DEMAC is intended to provide approximate results allowing the user to browse distributed data in a communication-efficient fashion. Then, once having identified a few patterns of interest, the relevant data can be centralized and precise analysis conducted. In our case, once the user has identified the approximate fundamental plane (88.6 percent of variance captured), the dataset can be downloaded to compute the precise variances.

To further illustrate the fundamental plane, we visualize the data projected onto the 1st and 2nd PCs, the 1st and 3rd PCs, the 2nd and 3rd PCs. Figure 5 displays the results. The left column depicts the projections onto the PCs computed by the centralized analysis. The right depicts the projections onto the PCs computed by our distributed algorithm at 50 percent communication.

From the left column figures, we see the fundamental plane. The top figure is the view perpendicular to the plane and the bottom two figures are the view perpendicular to the edge. From the right column, we see a similar pattern indicating that the distributed algorithm also allows identification of the plane.

---

5In terms of the similarity between its results and those of a centralized approach
6http://irsa.ipac.caltech.edu/applications/Gator/
8petroRad_i (galaxy view), z (SpecObj view) and velDisp (SpecObj view) in SDSS DR4
9\(K_{\text{msb}}\) in the extended source catalog in the All Sky Data Release, http://www.ipac.caltech.edu/2mass/release/allsky/index.html
10All of the preprocessing steps described above could have been carried out without any distributed computation, thus, need not enter into our simulation.
11All variances reported from the distributed algorithm (times 100) are the result of running the algorithm 100 times and averaging the results.
7 Conclusions

We proposed a system, DEMAC, for the distributed exploration of massive astronomical catalogs. DEMAC is to be built on top of the existing U.S. national virtual observatory environment and provide tools for data mining (as web services) without requiring datasets to be down-loaded to a centralized server. Instead, the users will only down-load the output of the data mining process (a data mining model); the actual data mining from multiple data servers will be performed using communication-efficient DDM algorithms. The distributed algorithms we have developed sacrifice perfect accuracy for communication savings. They offer approximate results at a considerably lower communication cost than that of exact results through centralization. As such, we see DEMAC as serving the role of an exploratory “browser”. Users can quickly get (generally quite accurate) results for their distributed queries at low communication cost. Armed with these results, users can focus in on a specific query or portion of the datasets, and down-load for more intricate analysis.

To illustrate the potential effectiveness of our system, we carried out a case study using distributed principal component analysis (PCA) for detecting fundamental planes of astronomical parameters. We observed our distributed algorithm to identify a fundamental plane (observed through centralized analysis) at reduced communication cost.

In closing, we envision our system to increase the ease with which large, geographically distributed astronomy catalogs can be explored, by providing quick, low-communication solutions. Such benefit will allow astronomers to better tap the riches of distributed virtual tables formed from joined and integrated sky survey catalogs.

Acknowledgments

H. Dutta, C. Giannella, H. Kargupta, R. Wolff: Support for this work was provided by the U.S. National Science Foundation (NSF) through grants IIS-0329143, IIS-009353, IIS-0201958 and by the U.S. National Aeronautics and Space Agency (NASA) through grant NAS2-37143.

K. Borne: Support for this work was provided in part by the NSF through Cooperative Agreement AST0122449 to the Johns Hopkins University.

References


Figure 5: Projections onto the PCs, communication percentage 50 percent.