SANDIA REPORT

SAND2011-4612
Unlimited Release
Printed July 2011

Topology for Statistical Modeling of Petascale Data

Janine Bennett, Philippe Pébay, David Thompson, Ajith Mascarenhas
Valerio Pascucci, Joshua Levine, Attila Gyulassy,
Maurice Rojas, and Korben Rusek

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation,
a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's
National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2011-4612
Unlimited Release
Printed July 2011

Topology for Statistical Modeling of Petascale Data

Janine Bennett, Philippe Pébay, David Thompson, Ajith Mascarenhas
Sandia National Laboratories
P.O. Box 969
Livermore, CA 94551, U.S.A.

Valerio Pascucci, Joshua Levine, Attila Gyulassy
Scientific Computing and Imaging Institute
University of Utah
72 So. Central Campus Drive, Room 3750
Salt Lake City, UT 84112, U.S.A.

Maurice Rojas, Korben Rusek
Department of Mathematics
Texas A&M University
TAMU 3368
College Station, TX 77843-3368, U.S.A.

Abstract

This document presents current technical progress and dissemination of results for the Mathematics for Analysis of Petascale Data (MAPD) project titled "Topology for Statistical Modeling of Petascale Data", funded by the Office of Science Advanced Scientific Computing Research (ASCR) Applied Math program.

Acknowledgments

The authors would like to thank Sandy Landsberg, DOE/ASCR Applied Mathematics program manager, for supporting this work.

Contents

1 Introduction . . . 7
2 Contributions in Combinatorial Topology . . . 9
2.1 Visualizing Morse Smale Complexes . . . 9
2.2 Unstructured Morse-Smale Complexes . . . 9
2.3 Edge Maps . . . 10
2.4 Quantized 2D Vector Fields . . . 12
2.5 Fiedler Trees for Multiscale Surface Analysis . . . 13
3 Contributions in Statistical Modeling . . . 15
3.1 Faster Real Solving . . . 15
3.2 Scalable Parallel Statistical Analysis . . . 24
4 New Integrated Topological and Statistical Methods . . . 31
4.1 Exploring High dimensional Spaces for Uncertainty Quantification . . . 31
4.2 Analysis of Large-Scale Scalar Data Using Hixels . . . 32
4.3 Feature-Based Statistical Analysis of Large Data . . . 34
5 Dissemination of results . . . 37
6 Planned Work . . . 39
References . . . 44

This page intentionally left blank

1 Introduction

Many commonly used algorithms for mathematical analysis do not scale well enough to accommodate the size or complexity of petascale data produced by computational simulations. The primary goal of this project is thus to develop new mathematical tools that address both the petascale size and uncertain nature of current data.

At a high level, our approach is based on the complementary techniques of combinatorial topology and statistical modeling. In particular, we use combinatorial topology to filter out spurious data that would otherwise skew statistical modeling techniques, and we employ advanced algorithms from algebraic statistics to efficiently find globally optimal fits to statistical models. This document summarizes the technical advances we have made to date that were made possible in whole or in part by MAPD funding. These technical contributions can be divided loosely into three categories: (1) advances in the field of combinatorial topology, (2) advances in statistical modeling, and (3) new integrated topological and statistical methods.

Roughly speaking, the division of labor between the three institutions (Sandia National Laboratories in Livermore, Texas A&M in College Station, and University of Utah in Salt Lake City) is as follows:

- The Sandia group focuses on statistical methods and their formulation in algebraic terms, and finds the application problems (and data sets) most relevant to this project.
- The group at University of Utah group develops new algorithms in computational topology via Discrete Morse Theory.
- The Texas A&M group develops new algebraic geometry algorithms, in particular with fewnomial theory.

In order to ensure a real synergy of ideas and convergence of efforts, three groups participating in this joint project remain in tight contact, in particular with bi-monthly video-conferences.

2 Contributions in Combinatorial Topology

2.1 Visualizing Morse Smale Complexes

Recent advances in practical algorithms for computing Morse-Smale (MS) complexes have made possible multi-resolution analysis of volumetric scalar valued data. Although these approaches are gaining popularity in analysis of scientific data, visualization techniques have not yet explored the full potential of this technology. In [GKK+, GBP, WBGP], we present novel visualizations using features extracted from MS complexes. We characterize possible visualizations enabled by the robust computation of all dimensional manifolds of the MS complex, and present several examples of these. Furthermore, we developed a framework for selecting features, assigning attributes, and building complex and compelling visualizations. In Figure 1 we show examples of topology-based techniques used to extract features that are hard to detect with more conventional methods not making use of topologic information.

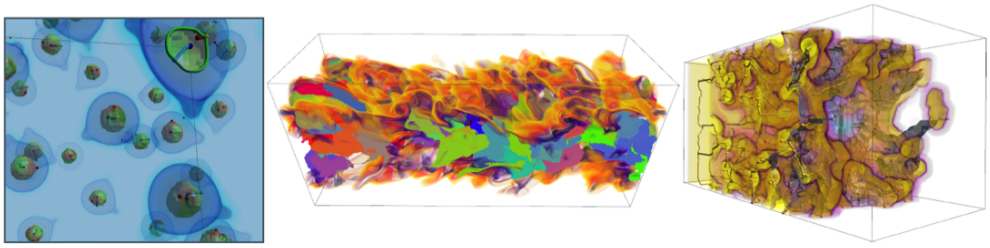
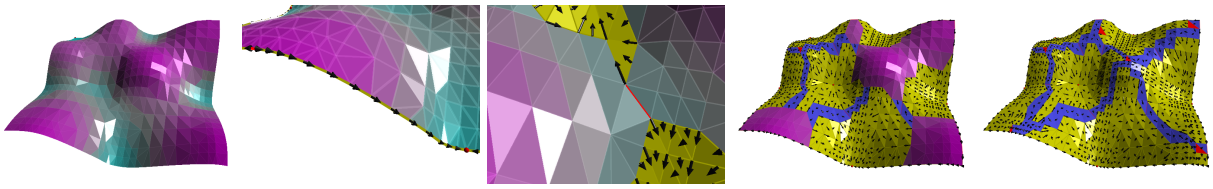


Figure 1. Left: Combinatorial computation of topological invariants results in robust identification of features, even in degenerate cases, such as topological pouches. Center: Using the machinery of topological persistence and simplification, we can visualize the 3-manifolds of the MS complex forming flow basins in a manner oblivious to noise. Right: Derived structures, for example, separating surfaces, can be used to represent non-physical phenomena, such as the "outer surface" of a sponge-like material.

2.2 Unstructured Morse-Smale Complexes

Given the success of topological analysis tools in many domains, there is a need to compute MS complexes for data that is defined on fully unstructured domains. Many of the first software milestones were designed for gridded domains. We have developed software tools which implement Gyulassy's discrete gradient construction [GBHP08] on regular cell complexes. This resulted in both a video tutorial as well as multimedia submission [GLP11]. The algorithm is illustrated in Figure 2. This generic software allows the construction of MS complexes for multiple modalities

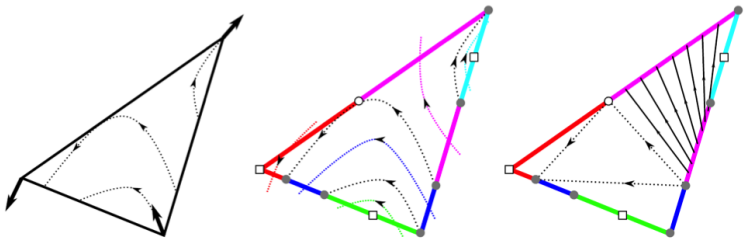
of data. As a result, it will enable topological analysis in for many data sources, allowing for the same demonstrated benefits we have already seen in gridded and piecewise-linear (PL) domains to be leveraged for multiple types of data.



(a) (b) (c) (d) (e)

Figure 2. Assigning gradient arrows on a terrain (a). Scalar values (height) are encoded from cyan (low values) to magenta (high values). (b) Boundary cells are paired first. (c) Pairing interior cells finds a saddle (red edge). (d) As pairing continues, a maxima is identified (red triangle). (e) Gradient construction is complete. Ascending 1-manifolds shown as blue cells.

2.3 Edge Maps



(a) (b) (c)

Figure 3. The structure of an edge map of a regular triangle. (a) The original triangle is represented as three vectors, which implies a flow throughout the interior. (b) Our representation subdivides the boundary into a set of intervals, which are broken at internal transition points (white circles), external transition points (white squares), and image points (grey circles). (c) Pairs of intervals are grouped into maps which represent sources and destinations of flow through the triangle intervals.

Robust analysis of vector fields has been established as an important tool for deriving insights from the complex systems these fields model. Traditional analysis and visualization techniques rely primarily on computing streamlines through numerical integration. The inherent numerical errors of such approaches are usually ignored, leading to inconsistencies that cause unreliable visualizations

and can ultimately prevent in-depth analysis. In [JBB+10, BJB+11a, JBB+11a, BJB+11b] we propose an alternate representation for vector fields on surfaces that explicitly represents the flow behavior of the field through each triangle. This representation, called *edge maps*, complements the traditional approach of storing sample vectors on the vertices of the triangulation. Figure 3 shows the structure of an edge map of a regular triangle.

One piece of this work [JBB+10, JBB+11a] focuses on the mathematical properties of edge maps. Edge maps allow for a multi-resolution approximation of flow by merging adjacent streamlines into an interval based mapping. Consistency is enforced at any resolution if the merged sets maintain an order-preserving property. At the coarsest resolution, we define a notion of equivalency between edge maps, and show that there exist 23 equivalence classes (Figure 4) describing all possible behaviors of piecewise linear flow within a triangle.

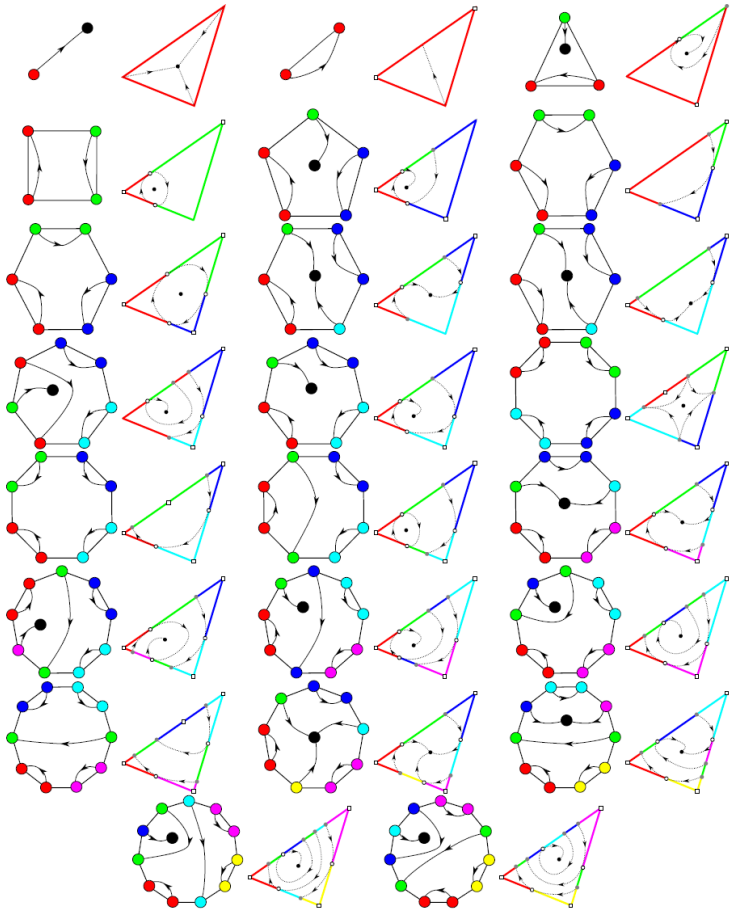


Figure 4. The 23 equivalent classes of mixed graphs that represent piecewise linear flow through a triangle, along with one possible rendition of the edge map. The ordering is of increasing number of links in the map.

A second branch of this work [BJB+11a, BJB+11b] focuses on encoding the spatial and temporal

errors which we use to produce more informative visualizations. This work describes the construction of edge maps, the error quantification, and a refinement procedure to adhere to a user defined error bound (Figure 5). Independent of this error all streamlines computed using edge maps are guaranteed to be consistent, enabling the stable extraction of features such as the topological skeleton We introduce new visualizations using the additional information provided by the edge maps to indicate the uncertainty involved in computing streamlines and topological structures (Figure 6). [BJB+11a] received a best paper award at the 4th IEEE Pacific Visualization Symposium in Hong Kong, China 2011.

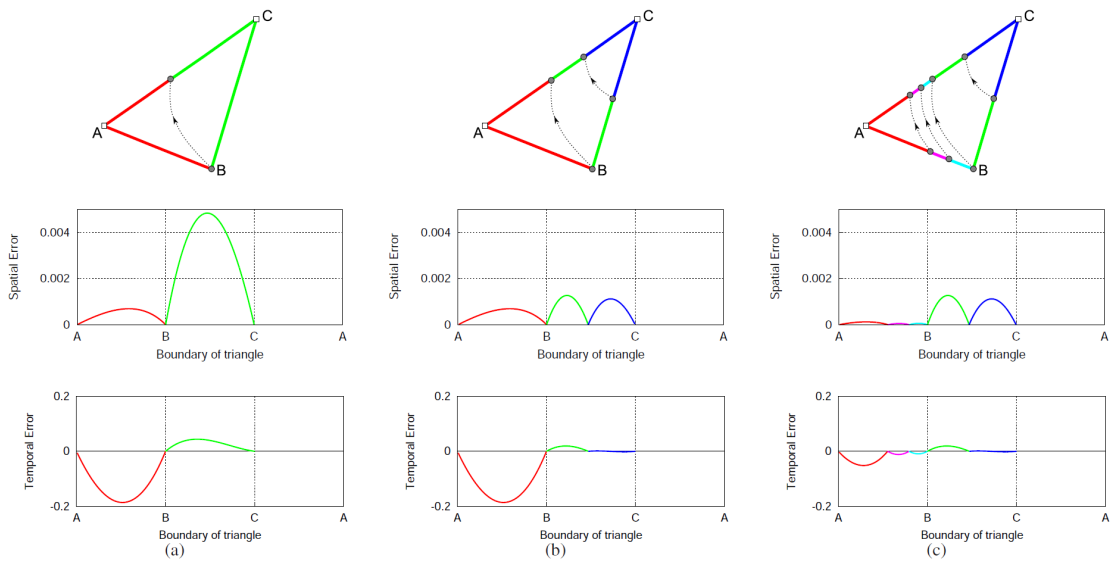


Figure 5. Reducing the mapping error (middle row: spatial error, bottom row: temporal error) by refinement of edge maps (top row). Level of refinement increases from left to right. The length of the edge AC is 0.0354, and the average time taken by a particle to travel across the triangle is 1.7. (a) No refinement. (b) Spatial refinement with an error bound of 0.003 splits the green link into two, creating two new links (green and blue) with smaller spatial and temporal errors. (c) Temporal refinement with an error bound of 0.06 splits the red link twice, creating three new links (red, cyan and magenta) with even smaller spatial and temporal errors.

2.4 Quantized 2D Vector Fields

Visualization and analysis of vector fields often hinges on the robust identification of structures such as critical points, separatrices, or closed orbits. Traditional techniques for computing these features fall broadly into two categories: (1) those that use numerical integration and (2) those that rely on purely combinatorial structures. However, the first set of tools often generate inconsistent

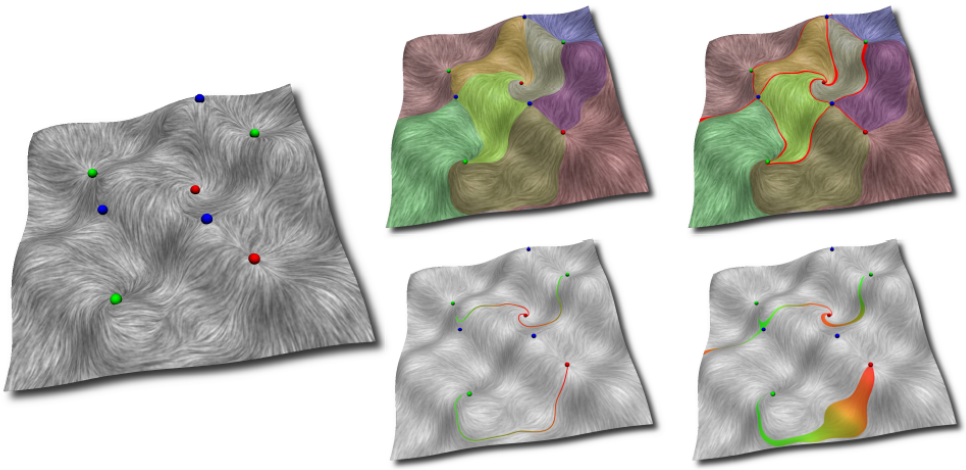


Figure 6. Edge maps enable new views of vector field stability, illustrated with a vector field on this wavy surface. Top row (middle right): A visualization of some colored regions where flow shares the same source (green spheres) and sink (red spheres) is augmented to show how these regions overlap when error is introduced. Bottom row (middle right): Streamwaves (colored green to red as they grow) show the advection of a single particle. In the presence of error, waves can widen and narrow, and bifurcate or merge.

results due to compounded approximation errors, while the second set often severely reduces the accuracy of the results. Instead, we propose [JBB+11b] a new discrete representation of vector fields that approximates the flow up to an arbitrary, user-defined error. By quantizing streamlines along edges of a triangulation we create a graph-based representation of the flow with up to 2^{32} nodes per edge. The graph is implicitly represented by a rasterization scheme and replaces streamline integration with a directed graph traversal. As a result, inherently unstable structures such as separatrices and cycles can be computed exactly and correctly up to the given approximation error. By varying the amount of discretization, we can provide a multi-resolution representation of vector fields that allows for a balance between storage space, computational effort, and fidelity to piecewise linear interpolation. Figure 7 shows our technique on ocean current data, illustrating a more complete topological representation that includes separatrices, stable manifolds, and cycles detected with our approach.

2.5 Fiedler Trees for Multiscale Surface Analysis

In [BNPS10] we introduce a new hierarchical decomposition method for multi-scale analysis of surface meshes. In contrast to other multi-resolution methods, our approach relies on spectral properties of the surface to build a binary hierarchical decomposition. Namely, we utilize the first

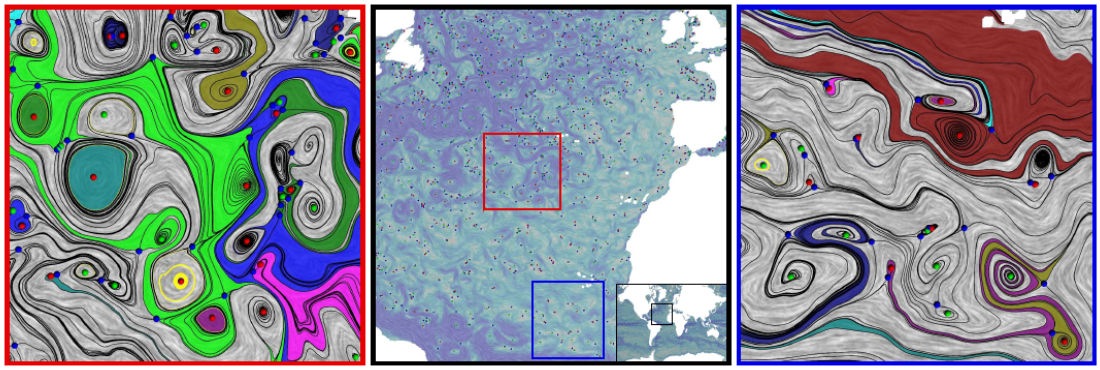


Figure 7. The oceanic currents of the North Atlantic ocean. In the center, we show a 600x600 vertex tile from a larger simulation of oceanic currents. Each image on the side is a zoomed in view visualizing the topology for the tile. Yellow lines are closed streamlines, the colored regions are stable manifolds grown from all the sinks (red balls), and the black lines are separatrices grown from all saddles (blue balls).

nontrivial eigenfunction (the Fiedler vector) of the Laplace-Beltrami operator to recursively decompose the surface. For this reason we coin our surface decomposition the Fiedler tree. Using the Fiedler tree ensures a number of attractive properties, including: mesh-independent decomposition, well-formed and nearly equi-areal surface patches, and noise robustness. We show how the evenly distributed patches can be exploited for generating multi-resolution high quality uniform meshes. Additionally, our decomposition permits a natural means for carrying out wavelet methods, resulting in an intuitive method for producing feature sensitive meshes at multiple scales. The method and its generalization to volumetric meshes will be a critical component in building the hierarchical models needed in analysis of petascale data. Figure 8 shows the intrinsic nature of the mesh decomposition.

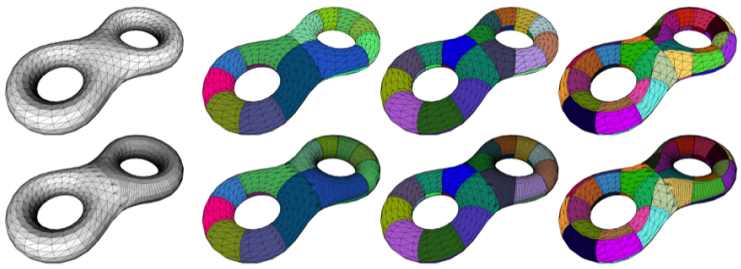


Figure 8. Two input meshes with same shape and different triangulation (left) yield the same sequence of decompositions (in color).

3 Contributions in Statistical Modeling

3.1 Faster Real Solving

The complexity of polynomial system solving is important because of its ancient roots, its numerous connections to geometry and number theory, and the need for reliable large-scale numerical computation, e.g., the problems arising from this project where we must fit petascale data to statistical models. We thus henceforth focus on the numerical approximation of real solutions of large systems of polynomial equations. Our key advance is a method to solve polynomial systems with few real solutions and many complex solutions, using an effort **logarithmic** in the number of complex solutions (Theorem 3.1 below). The best previous methods have complexity polynomial (or worse) in the number of complex solutions.

Definition 3.1. We call an $f \in \mathbb{R}[x_1,\ldots,x_n]$ (with $f(x)=\sum_{i=1}^{n+k} c_i x^{a_i}$, $c_i \neq 0$ and $x^{a_i}=x_1^{a_{1,i}}\cdots x_n^{a_{n,i}}$ for all i, and the a_i distinct) an ***n*-variate $(n+k)$-nomial**. We also define $\mathrm{Supp}(f):=\{a_1,\ldots,a_{n+k}\}$ to be the **support** of f. We denote the collection of n-variate $(n+k)$-nomials in $\mathbb{R}[x_1,\ldots,x_n]$ by $\mathscr{F}_{n,n+k}$. Finally, if $F:=(f_1,\ldots,f_n)$ with $f_i \in \mathscr{F}_{n,n+k}$ and $\mathrm{Supp}(f_i)=\{a_1,\ldots,a_{n+k}\}$ for all i then we call F a **(real) $(n+k)$-sparse $n \times n$ polynomial system**. $\diamond$

While our methods are completely general, a small example will help illustrate our central ideas.

Example 3.1. Let us start with the 6-sparse 2×2 polynomial system

$$F := \begin{cases} x_1^{82} + \frac{31}{50}x_2^{41} - x_2 \\ x_2^{82} + 55x_1^{41} - x_1 \end{cases}$$

The classical Bézout's Theorem [Har] implies that F has no more than $82^2 = 6724$ complex roots (assuming F has only finitely many). A more recent **fewnomial bound** of Li, Rojas, and Wang [LRW03, Thm. 1] tells us that F has no more than 5 roots in the positive quadrant $\mathbb{R}^2_+$ (assuming F has only finitely many). Counting the exact number of real roots for our example turns out to be more subtle than one may expect: the usual methods of Gröbner bases or resultants (on the well-known computer algebra systems `Maple 13`, `Singular`, and `Macaulay2`) result in "out of memory" errors within either a few minutes or a few hours. (We briefly review these techniques in the next section.) On the other hand, via a preliminary `Matlab` implementation of our techniques here, we can determine within a few seconds that F has exactly 1 (resp. 2, 2, 0) root(s) in $\mathbb{R}^2_+$ (resp. $\mathbb{R}_- \times \mathbb{R}_+$, $\mathbb{R}^2_-$, $\mathbb{R}_+ \times \mathbb{R}_-$). $\diamond$

Regardless of what polynomial system solving technique one ultimately favors, one is inevitably led to consider certain intricate regions within certain families of polynomial systems. More precisely, let us now define

$$F_{(a,b)} := \begin{cases} x_1^{82} + a x_2^{41} - x_2 \\ x_2^{82} + b x_1^{41} - x_1 \end{cases}$$

and let ∇ denote the closure of the set of all $(a,b) \in \mathbb{C}^2$ resulting in an $F_{(a,b)}$ with degenerate roots. Our set ∇ is an example of a **discriminant variety** (or a collection of **ill-posed** problems). Here, ∇ is a curve in $\mathbb{C}^2$, and the complement of the **real part** of ∇ in $\mathbb{R}^2$ determines connected regions (known as **chambers**) on which the number of real roots of $F_{(a,b)}$ is constant. (This is detailed rigorously, and in greater generality, in the next section.) In particular, determining the exact number of real roots of F can be reduced to determining the chamber containing F. More to the point, it has long been known in numerical linear algebra and optimization that the complexity of numerical problems depends critically on the distance to ill-posedness, i.e., how far one is from an underlying discriminant variety.

Theorem 3.1. *Fix n and let $\mathscr{A} = \{a_i\} \subset \mathbb{Z}^n$ have cardinality $n+k$. Then, in time polynomial in the sparse encoding, we can determine the unique chamber cone (for the $\mathscr{A}$-discriminant amoeba) containing $f(x) = \sum_{i=1}^{n+k} c_i x^{a_i}$, or obtain a true declaration that f lies in ≥ 2 chamber cones.* $\square$

While the defining polynomial Δ for our example discriminant ∇ has coefficients with over 6000 digits, we can nevertheless make considerable practical use of ∇. We describe later how we can circumvent such bottlenecks.

A consequence of our notion of input size is that we have complexity polynomial in the **logarithm** of the degree — a tremendous speed-up over earlier computational algebra methods. We conjecture that a similar speed-up holds when we fix k and let n grow instead. We also point out that while Theorem 3.1 does not appear to involve polynomial systems, the **Cayley Trick** [GKZ94] is a simple construction that allows one to apply the theorem above to any $\ell \times n$ system of equations with $\ell \leq n$. (Indeed, our preceding construction of ∇ used the Cayley Trick, ultimately employing a 6-variate trinomial.)

Our theorem above also refines an earlier tropical result where polynomial complexity is proved for a similar membership problem involving a simpler complex of cones intersecting at the origin [DFS07]. (Technically, the polyhedral complex defined by our chamber cones is more complicated because our cones do **not** intersect at the origin.) A key trick also not present in earlier tropical geometric or computational algebraic work is to observe that deciding chamber cone membership involves checking the sign of a linear combination of logarithms. So one needs to avail to Baker's famous theorem on linear forms in logarithms [Nes03].

Some Additional Background

The **Horn-Kapranov Uniformization** [Kap91, PT05] yields, quite surprisingly, a succinct **one-line rational parametrization** of any $\mathscr{A}$-discriminant variety. Applied to our curve ∇ (which has an unwieldy defining polynomial Δ), the Horn-Kapranov Uniformization results in a map $\varphi : \mathbb{P}^1_{\mathbb{C}} \times (\mathbb{C}^*)^4 \longrightarrow \mathbb{C}^6$ defined explicitly via:

$$\varphi(\lambda, t) := [\lambda_1, \lambda_2] \begin{bmatrix} -40 & 6723 & -6683 & -3280 & 0 & 3280 \\ -40 & 163 & -123 & -80 & 80 & 0 \end{bmatrix} \odot \left(1, \frac{t_2^{41}}{t_1^{82}}, \frac{t_2}{t_1^{82}}, \frac{t_2^{82} t_3}{t_1^{82}}, \frac{t_3}{t_1^{41}}, \frac{t_3}{t_1^{81}}\right).$$

Quotienting out by certain natural homogeneities then results in a rational map (composed with radicals) $\tilde{\varphi} : \mathbb{P}^1_{\mathbb{C}} \longrightarrow \mathbb{C}^2$ with $\tilde{\varphi}(\mathbb{P}^1_{\mathbb{C}}) = \nabla$. Letting $\mathrm{Log} : \mathbb{C}^n \longrightarrow \mathbb{R}^n$ denote the map defined by $(\log|x_1|, \ldots, \log|x_n|)$, we define the image of the complex zero set of any polynomial g under Log to be the **amoeba** of g, written $\mathrm{Amoeba}(g)$. Using $\tilde{\varphi}$ we can then easily plot $\mathrm{Amoeba}(\Delta)$, and observe that the complement of $\mathrm{Amoeba}(\Delta)$ appears to consist of a finite union of convex sets. This is in fact a special case of a more general theorem on amoeba complements [GKZ94]. What is most remarkable, however, is that the map $\tilde{\varphi}$ allows us to easily plot $\mathrm{Amoeba}(\Delta)$ even when the monomial term expansion of Δ is beyond the range of any current computational algebra software.

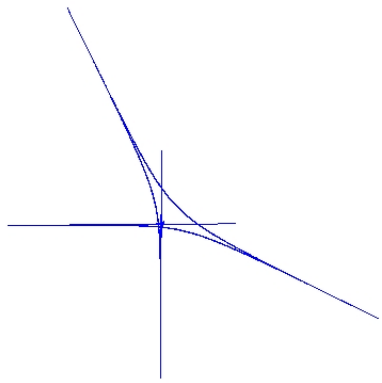


Figure 9. Connected subsets of the chambers of corresponding to the discriminant variety ∇.

For instance, as illustrated in Figure 9, we can restrict our parametrization $\tilde{\varphi}$ to $\mathbb{P}^1_{\mathbf{R}}$: the connected components are actually images of connected subsets of the chambers of our discriminant variety ∇. We will then abuse notation slightly by referring to the unbounded connected components of the complement of $\mathrm{Amoeba}(g)$ as **outer chambers**. The remaining connected components of the complement of the contour of $\mathrm{Amoeba}(g)$ are called **inner chambers**.

Remark 3.1. The key observation at this stage is that (a) it appears to be more likely to lie in an outer chamber than an inner chamber, and (b) the number of real roots of $F_{(a,b)}$ is constant for (a,b) in a fixed (inner or outer) chamber. $\diamond$

The outer chambers of ∇ in fact correspond to **mixed subdivisions** of the pair of supports coming from F. Mixed subdivisions are a type of polyhedral complex that is essentially a triangulation of a Minkowski sum of point sets, endowed with additional structure [HS95]. That outer chambers of discriminant amoebae correspond to mixed subdivisions can then be derived via the theory in [GKZ94].

The precise construction of these subdivisions need not concern us now, but a consequence of the underlying theory is that one can associate a collection of **lower binomial systems** to each such mixed subdivision. In particular, via a method we will describe shortly, we can determine that $\left(-\log\frac{31}{50}, -\log 55\right)$ lies exactly in the outer chamber of Figure 10, indicated by the single remaining mixed subdivision.

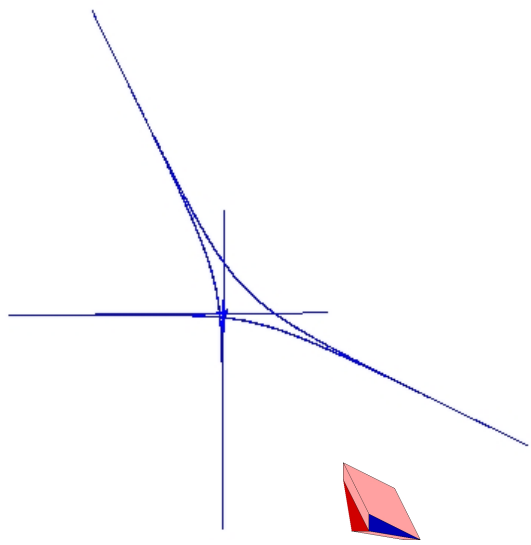


Figure 10. Connected subsets of the chambers of corresponding to the discriminant variety ∇, and mixed subdivision for $\left(-\log\frac{31}{50}, -\log 55\right)$.

For the indicated outer chamber and mixed subdivision, the resulting collection of lower binomial systems for F is exactly the following:

$$x_1^{82} + \tfrac{31}{50}x_2^{41} \qquad \tfrac{31}{50}x_2^{41} - x_2 \qquad x_1^{82} - x_2$$

$$x_2^{82} + 55x_1^{41} \qquad x_2^{82} + 55x_1^{41} \qquad 55x_1^{41} - x_1$$

Writing 4-tuples for the number of roots in $\mathbb{R}_+^2$, $\mathbb{R}_- \times \mathbb{R}_+$, $\mathbb{R}_-^2$, and $\mathbb{R}_+ \times \mathbb{R}_-$, one can then easily check that our 3 binomial systems respectively have the following distributions of roots in $(\mathbb{R}^*)^2$: $(0,0,1,0)$, $(0,1,1,0)$, and $(1,1,0,0)$. In other words, we have just shown that the distribution of roots for F is $(1,2,2,0)$ as promised.

Even better, the construction of our mixed subdivisions also yields an explicit **homotopy** (or **toric deformation**) that allows us to efficiently approximate the roots. To be precise, the construction of the mixed subdivision above entails **lifting** the underlying support points and then finding certain edge pairs from the lifted point sets [HS95]. The lifting values for our running example yield the following **lifted** system with an extra parameter:

$$\widehat{F_t} := \begin{cases} x_1^{82} + \frac{31}{50}\underline{\boldsymbol{t}^{\mathbf{1}}}x_2^{41} - x_2 \\ x_2^{82} + 55\underline{\boldsymbol{t}^{-\mathbf{8}}}x_1^{41} - x_1 \end{cases}$$

(In particular, this system corresponds to assigning height 0 to all the points, save for a height of 1 for $(0,41)$ and a height of -8 for $(41,0)$.) We then re-tailor the method of **polyhedral homotopy** (normally used to find **all** complex solutions) as follows: use just the **real** roots of the lower binomial systems as initial guesses for the real roots of F. Then, using standard numerical

continuation with t going from 0 to 1, deform the initial guesses into approximations of the roots of F.

Remark 3.2. Our method thus complements the known technique of polyhedral homotopy by providing **canonical** liftings for a given F. Indeed, to this day, polyhedral homotopy methods always employed random liftings. Furthermore, by restricting to just the real roots of the lower binomial systems, we obtain a simple and substantial computational speed-up. $\diamond$

The resulting speed-ups are significant, even when applied in a preliminary way to numerical code written before this project. For example, using Jan Verschelde's `PHC` code, the lifting found by knowledge of the outer chamber containing F allows us to numerically solve F (and detect our 5 real roots in the correct quadrants) within 10 seconds. In contrast, allowing random liftings results in a slow-down by a factor of 2.8, and frequent miscounting of (or complete failure to find) real roots. We also point out that T. Y. Li's `HOM4PS2` code (slightly modified to allow user-specified liftings) successfully detected the real roots of F (in their correct quadrants) within 0.87 seconds, apparently independent of the chosen lifting. (There was a mild speed-up, under 1%, when `HOM4PS2` was given the lifting corresponding to the chamber containing F.) However, both software packages worked with **all** 6724 complex solution paths, so there is still a significant speed-up that we have not yet tapped: following just the 5 **real** homotopy paths. We are currently working on incorporating the latter speed-up as well.

Remark 3.3. Finding homotopy algorithms preserving the number of real roots was an open problem. Part of our work on this proposal thus provides a solution. $\diamond$

Algorithm in detail

To illustrate the algorithm we have derived during Year 2, we will discuss a particular family of 3×3 polynomials as a running example.

Example 3.2. Consider the family of polynomial system obtained by setting certain coefficients to zero in a 3×3 unmixed 9-nomial system:

$$F := \begin{cases} x^6 + ay^3 - z \\ y^6 + bz^3 - x \\ z^6 + bx^3 - y \end{cases}$$

where a, b, c are real constants to be specialized later. The classical Bézout's Theorem [Har] implies that F has no more than $6^2 = 216$ complex roots, assuming F has only finitely many. However, the number of positive roots is, with high probability, no greater than 2. Moreover, we can efficiently determine the exact number of positive roots (given real a, b, and c) via linear programming. How we accomplish the last two claims is described below. $\diamond$

Using our example above, let us now summarize the general technique we used to determine the sharper count for the number of positive roots:

1. Form a single lifted support set $\widehat{\mathscr{A}}$ via the Cayley Trick: For our preceding example, this means we replace the 3 sets of exponent vectors $\{(6,0,0),(0,3,0),(0,0,1)\}$, $\{(6,0,0),(0,3,0),(0,0,1)\}$, and $\{(6,0,0),(0,3,0),(0,0,1)\}$ by the following single matrix:

$$\widehat{\mathscr{A}} := \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 3 & 0 & 6 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 3 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

2. Construct the Gale Dual of $\mathscr{A}$: Letting $U\widehat{\mathscr{A}}^{\top} = H$ denote the Hermite Factorization [Sto00, Ch. 6, Table 6.2, pg. 94] of the transpose of $\widehat{\mathscr{A}}$, suppose $\widehat{\mathscr{A}}$ has dimensions $(n+1)\times(n+k)$ and consider the matrix B that is the transpose of the bottom $k-1$ rows of U. For instance, in our running example, we obtain this matrix:

$$B := \begin{bmatrix} 3 & -1 & 0 \\ -6 & 1 & -6 \\ 3 & 0 & 6 \\ 2 & 0 & 3 \\ 1 & -6 & 0 \\ -3 & 6 & -3 \\ -1 & 3 & -1 \\ -5 & 0 & 1 \\ 6 & -3 & 0 \end{bmatrix}.$$

3. Define a hyperplane arrangement from certain rows of B Suppose the rows of B are $b_1,\ldots,b_{n+k}$. Call *radiant* any set of indices $\mathscr{I}\subset\{1,\ldots,n+k\}$ satisfying the two conditions:

(a) $[b_i]_{i\in\mathscr{I}}$ is a maximal rank 1 sub-matrix of B,

(b) $\sum_{i\in\mathscr{I}} b_i$ is *not* the zero vector.

We then let $\mathscr{H}\subset\mathbb{P}^{k-2}_{\mathbf{R}}$ denote the union of hyperplanes that are perpendicular to the line generated by $\sum_{i\in\mathscr{I}} b_i$ for some radiant subset $\mathscr{I}$. For our example, $\mathscr{H}$ is the arrangement of 9 green lines to the right; and there are exactly 9 radiant subsets, each with cardinality 1. One then needs to find every index set $\mathscr{J}$ of a set of hyperplanes defining an angle cone of some vertex of $\mathscr{H}$.

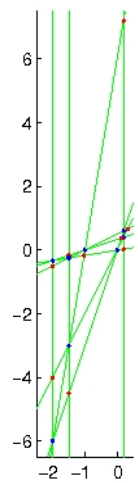


Figure 11. The set $\mathscr{J}$ corresponding to Example 3.2. Some lines are close together.

For instance, one can see with our running example 17 vertices, and a triplet of parallel lines yielding an $18^{\underline{\text{th}}}$ vertex at infinity, as illustrated in Figure 11. One can also see that there are 33 such sets $\mathscr{J}$ visible from the illustration, and another 3 coming from a triple intersection of lines at infinity. More precisely, each red (resp. blue) vertex defines one (resp. a triplet of) such index set(s).

4. Collections of hyperplanes at an angle cone generate new cones For each $\mathscr{J}$ found in the last step, we obtain the $(k-2)$-dimensional cone $W_{\mathscr{J}}$ generated by the vectors $\{-b_j\}_{j\in\mathscr{J}}$. The union of these cones then defines a new polyhedral complex $\mathscr{C}'$ in $\mathbb{P}^{k-2}_{\mathbf{R}}$ which we call a *(pointed) cone* arrangement. For our example, our 36 cones intersect to form exactly 53 top-dimensional cells in $\mathscr{C}'$.

5. Shift the walls to get chamber cones For any $t\geq 0$ and $y=(y_1,\ldots,y_N)\in\mathbb{C}^N$ let us define $\log^+ t$ to be 0 or $\log t$ according as t is 0 or not; and set $\log^+|y|:=(\log^+|y_1|,\ldots,\log^+|y_N|)$. We then define a new (non-pointed) cone arrangement $\mathscr{C}$ from $\mathscr{C}'$ by shifting the facets of top-dimensional cells. In particular, expressing each top-dimensional cell $\sigma'\in\mathscr{C}'$ in the form $\bigcap_{\mathscr{J}} H_{\mathscr{J}}$ where $H_{\mathscr{J}}$ is the

half-space uniquely determined by the index set $\mathscr{J}$ and the cell σ', we define the (not necessarily pointed) cone $\sigma := \bigcap_{\mathscr{J}} (H_{\mathscr{J}} - \log^+ |v_{\mathscr{J}} B^T| B)$. $\mathscr{C}$ is then the polyhedral complex defined by the intersection of all such σ. (Note that while every cell of $\mathscr{C}'$ is unbounded in $\mathbb{R}^{k-1}$, $\mathscr{C}$ may possess bounded cells.) We call the co-dimension 1 cells of $\mathscr{C}$ *walls*. We also call the top-dimensional cells of $\mathscr{C}$ intersecting infinity *chamber cones*. The chamber cones of $\mathscr{C}$ are in bijective correspondence with the top-dimensional cells of $\mathscr{C}'$. Figure 12 illustrates our example.

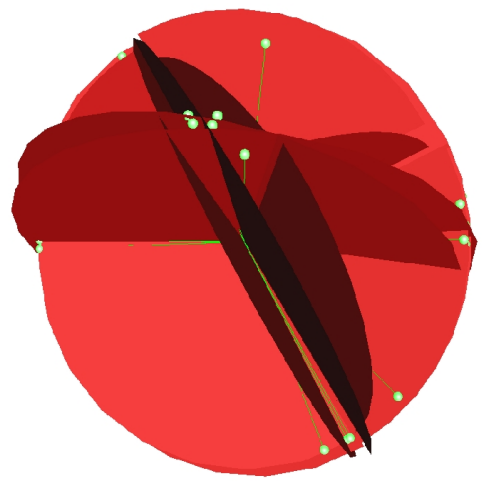


Figure 12. Example of chamber cones, limited by walls. Note that the "inner" cells are obscured.

6. Final preprocessing for fast counting Via the computational geometry technique of ε-cuttings [Cha01] (a higher-dimensional analogue of sorting), we build a data structure that allows us to decide which chamber cone contains a given polynomial system within a number of arithmetic operations polynomial in n and k.

Classically, Morse theory tells us that the number of roots of F is constant within any $\mathscr{A}$-discriminant chamber (see, e.g., [GKZ94]). So counting the number of real solutions of a polynomial system F is can be accomplished by deciding which chamber contains F (and knowing the correct count for the chamber).

The chamber cones we have constructed are in fact "outer" approximations of certain $\mathscr{A}$-discriminant chambers. Each chamber cone contains a unique $\mathscr{A}$-discriminant chamber, and the corresponding chamber is called an *outer* chamber. In particular, within each outer chamber, combinatorics rules: the number of positive roots is exactly the number of alternating mixed cells in a mixed subdivision uniquely defined by the corresponding chamber cone.

The case of Example 3.2 is illustrated in Figure 13, with a 2-dimensional slice of the discriminant variety (left) and the mixed subdivisions (right). In particular, no mixed subdivision has more than 2 mixed cells, thus implying that no F in an outer chamber has more than 2 positive roots.

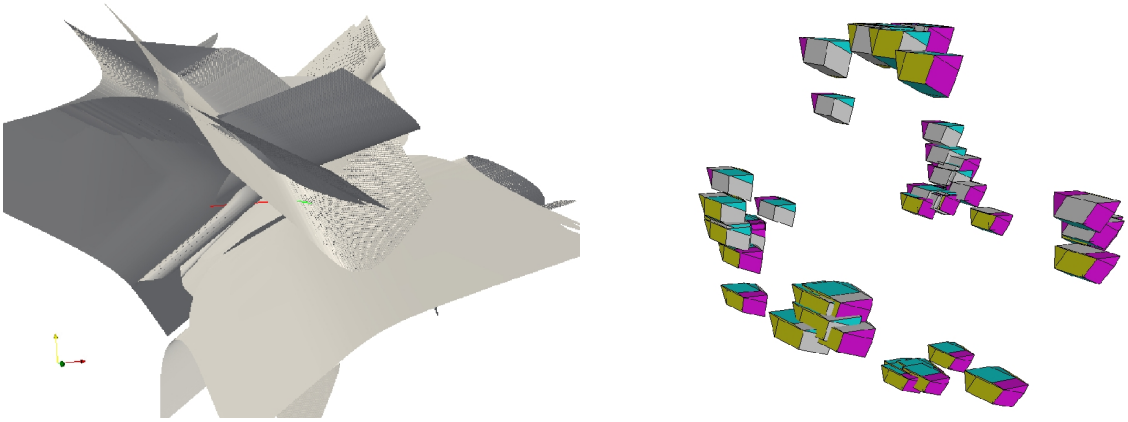


Figure 13. A 2-dimensional slice of the discriminant variety (left) and mixed subdivisions (right) for Example 3.2.

Our technique for fast real root counting is then to determine which outer chamber contains our given polynomial system F. Furthermore, the corresponding mixed subdivision contains additional data we can use to efficiently numerically approximate the real roots of F. (Computing mixed cells can in fact be embedded in our earlier pre-processing, and is another standard computational geometry task.) With respect to certain probability measures, random polynomial system lie in outer chambers with high probability: our recent paper [BHPR11] proves this in an important family of examples.

Remark 3.4. We emphasize that the 6 steps we have just described are *preprocessing that need only be done once per set of supports*. Once the resulting data structure is in place, querying how many real roots a system has is extremely efficient. $\diamond$

Much of our work this year consisted in pinning down the preceding structure of our algorithm, and implementing it in various special cases. In particular, we have implemented the cases $k \leq 3$ (with n arbitrary) in `Matlab`, and are currently completing a version in `Sage`. The latter version will be useful for further open-source dissemination and error-checking against our earlier version.

New Directions

The theory of chamber cones is actually closely related to certain investigations into algorithmic arithmetic geometry by Rojas. In particular, the papers [nIRR10, nIRRar] (which deal with polynomials over the p-adic rational numbers) arose from this work, and in turn helped clarify the structure of our algorithms over $\mathbb{R}$.

Rojas' Ph.D. student Rusek has also extended the quantitative aspects of fewnomial theory in a new direction: polynomials supported on structured point sets. In particular, he has succeeded in considerably sharpening Khovanski's famous upper bounds on the number of real roots for certain

specially structured polynomial systems [RSST].

Closer to real polynomial system solving, Rojas has also investigated alternative strategies from semidefinite programming. In particular, with Ph.D. student Rohun Kshirsagar, Rojas investigated the fraction of nonnegative polynomials expressible as sums of squares of polynomials. (The latter polynomials are called *sos*.) The motivation is that semidefinite programming can be used to efficiently optimize polynomials that are sos. In particular, should most nonnegative polynomials be sos, semidefinite programming would then be applicable to most polynomials. (It should be mentioned that Hilbert's 17th Problem concerned the expression of nonnegative polynomials as sums of squares of *rational functions*.)

It is known from work of Blekherman [Ble06] that for fixed degree, the fraction of nonnegative polynomials that are sos tends to 0 as the number of variables goes to infinity — a negative result. Rojas' investigation with Kshirsagar concerns a different but practically important setting: fixing the number of variables. During Year 2, Rojas and Kshirsagar found a Markov chain method to compute this fraction for low degree and a small number of variables (a setting which is still completely unexplored). This will lead to a new paper in Year 3.

Finally, we mention that Hauenstein has continued advancing his `C` code for numerical homotopy and, through the funding from this project, he has written a new paper on approximating points on real algebraic hypersurfaces [Hau11].

3.2 Scalable Parallel Statistical Analysis

Design Trade-Offs and Limiting Cases for Computing Quanta-Based Statistics in Parallel

Statistical analysis is typically used to reduce the dimensionality of and infer meaning from data. A key challenge of any statistical analysis package aimed at large-scale, distributed data is to address the orthogonal issues of parallel scalability and numerical stability. Many statistical techniques, e.g., descriptive statistics or principal component analysis, are based on moments and co-moments and, using robust online update formulas, can be computed in an embarrassingly parallel manner, amenable to a map-reduce style implementation.

In [PTB10] we focus on contingency statistics, one case of the more general class of quanta-based statistical methods. By the means of a primary models consisting of a contingency tables, several derived quantities, such as joint and marginal probability, point-wise mutual information, information entropy, and χ^2 independence statistics can be directly obtained. However, contingency tables can become large as data size increases, requiring a correspondingly large amount of communication between processors. This potential increase in communication prevents optimal parallel speed-up and is the main difference with moment-based statistics, where the amount of inter-processor communication is independent of data size. We therefore present the design trade-offs which we made to implement the computation of contingency tables in parallel and study the parallel speed-up and scalability properties of our open source implementation. In particular, we observe optimal speed-up and scalability when the contingency statistics are used in their appropri-

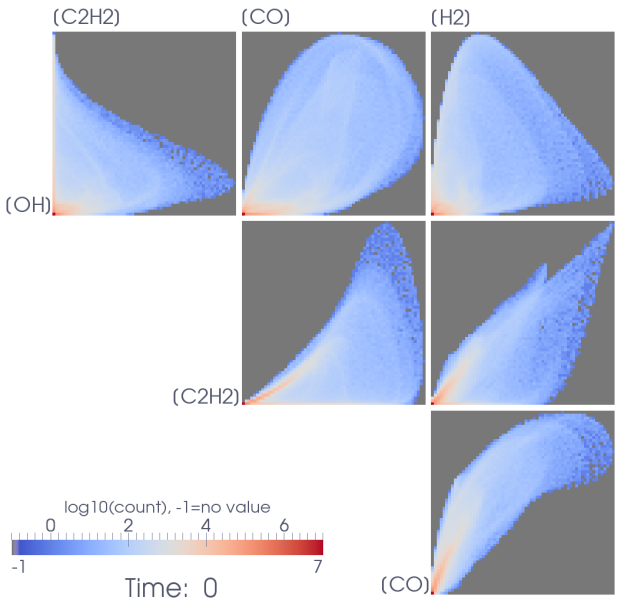


Figure 14. A set of 6 two-way contingency tables marginalized from a single four-way contingency table, taken from simulated combustion of a lifted ethylene jet. Plots in each row have an x axis showing concentrations of the chemical species to the left and plots in each column have a y axis showing concentrations of the species to the top. Color is used to indicate the number of times each set of concentrations were observed simultaneously at one point in the simulation domain. Only 4 of the 21 different species concentrations tracked were used to build the four-way table.

ate context, namely, when the data input is not quasi-diffuse. Figure 14 shows a set of 6 two-way contingency tables, marginalized from a single four-way contingency table, taken from a simulated combustion of a lifted ethylene jet.

In [PTBM11] we extend and generalize these results for another case of quanta-based statistics, namely, order statistics, where the primary model is represented by histogram, whereby the input data is also quantized. In particular, we establish that when used in their proper context, that is, when the input data is honestly discrete (as opposed to quasi-diffuse), then our parallel design trade-offs allow for near-optimal scalability, demonstrated with up to 10,000 processes of `jaguar`, the premier computing facility of DOE, as illustrated in Figure 15.

Design and Performance of a Scalable, Parallel Statistics Toolkit

Most statistical software packages implement a broad range of techniques but do so in an ad hoc fashion, leaving users who do not have a broad knowledge of statistics at a disadvantage since

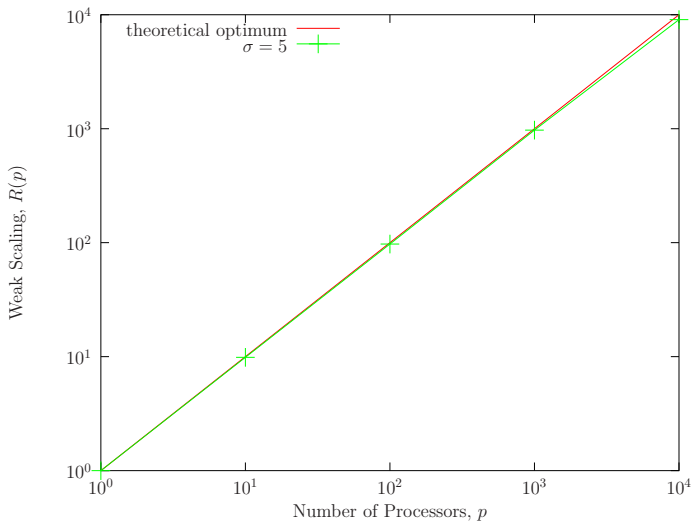


Figure 15. Parallel order statistics: weak scaling (at constant work per process) on `jaguar`, with $N(p)/p = 2 \times 10^7$.

they may not understand all the implications of a given analysis or how to test the validity of results. These packages are also largely serial in nature, or target multicore architectures instead of distributed-memory systems, or provide only a small number of statistics in parallel.

In [PTBM11] we survey a collection of statistics algorithm implementations developed as part of a common framework that groups modeling techniques with associated verification and validation techniques to make the underlying assumptions of the statistics more clear. Furthermore it employs a design pattern specifically targeted for distributed-memory parallelism, where architectural advances in large-scale high-performance computing have been focused. Specifically, we partition the statistical analysis workflow into 4 operations: Learn a model from observations, Derive statistics from a model, Assess observations with a model, and Test the null hypothesis. Figure 16 contains an example that illustrates the parallel execution of Learn, Derive and Assess operations of order statistics on 3 processes.

This parallel toolkit is released as open-source software (BSD-style license), as part of VTK [Kit10], itself part of the Titan Informatics Toolkit [Ttn] developed jointly by Sandia and Kitware. It is publicly available with Git; detailed instructions are provided by this webpage:

`http://www.vtk.org/Wiki/VTK/Git`

At the time of writing, the following 7 parallel engines workflows are implemented:

- descriptive statistics,
- histograms and order statistics,

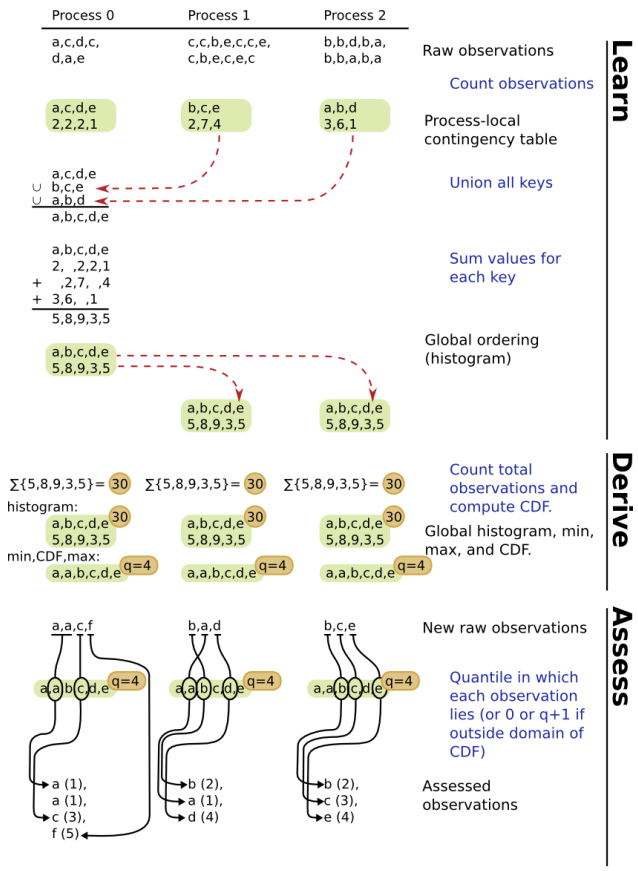


Figure 16. Example showing parallel execution of the Learn, Derive, and Assess operations of order statistics on 3 processes.

- bivariate linear correlation and regression,
- contingency statistics and information entropy,
- multi-variate linear correlation,
- principal component analysis,
- *k*-means clustering.

The corresponding operations provided by each of these 7 engines are detailed in Table 1.

A number of programs which make use of the parallel statistics classes of Titan are available in the `VTK/Infovis/Testing/Cxx/` sub-directory of VTK and, as such, belong to the test harness of VTK which is built and tested nightly on several systems, both serial and parallel. At the time of writing, more than thirty different platforms, with various operating and scheduling systems, are building the toolkit, running its test harness, and reporting the corresponding results to the Kitware Dashboard each night at:

http://www.cdash.org/CDash/

Moreover, most of the functionality of this toolkit is now available through the graphical user interface of the ParaView parallel visualization application [Hen04], also a joint effort between Sandia and Kitware, an application utilized by several thousands of users worldwide for scientific visualization and data analysis.

Table 1. The different operations currently made available by the parallel statistics engines.

	Learn	Derive	Assess	Test
Descriptive	Calculate minimum, maximum, mean, and centered M_2, M_3 and M_4 aggregates [BPRT09]	Calculate variance, standard deviation, skewness, and kurtosis (various estimators available for each statistic)	Mark with relative deviation (one-dimensional Mahalanobis distance [Mah36])	Calculate Jarque-Bera statistic [JB87] and perform χ^2 goodness of fit test
Order	Calculate histogram	Calculate arbitrary quantiles (e.g., quartiles, deciles, etc.)	Mark with quantile index	Calculate Kolomogorov-Smirnov test statistic [DCD86]
Correlative	Calculate minima, maxima, means, and centered M_2 aggregates [BPRT09]	Calculate variances, covariance, Pearson correlation r, and both linear regressions	Mark with squared two-dimensional Mahalanobis distance [Mah36]	Calculate bivariate Jarque-Bera-Srivastava statistic [KOS09] and χ^2 goodness of fit test
Contingency	Calculate the bivariate contingency table (also called a 2-dimensional histogram)	Calculate joint, conditional, and marginal probabilities, as well as information entropies	Mark with joint and conditional PDF values, as well as pointwise mutual information	Calculate Pearson χ^2 test of independence without and with Yates correction [Yat34]
Multi-Correlative	Calculate means and pairwise centered M_2 aggregates [BPRT09]	Calculate covariance matrix and its (lower) Cholesky decomposition	Mark with squared multi-dimensional Mahalanobis distance [Mah36]	N/A
PCA	Identical to the multi-correlative algorithm	Identical to the multi-correlative algorithm, plus the eigenvalues and eigenvectors of the covariance matrix [LV07]	Mark with coordinates in basis of all, or only first eigenvectors with cumulative energy above a given threshold	Calculate multivariate Jarque-Bera-Srivastava statistic [KOS09] and perform χ^2 goodness of fit test
k-Means	Compute k cluster centers given a positive integer k [Mac67]	Calculate global and local rankings amongst sets of clusters, and total error [BPT09]	Mark with closest cluster id and associated distance for each set of cluster centers	In progress

4 New Integrated Topological and Statistical Methods

4.1 Exploring High dimensional Spaces for Uncertainty Quantification

An important goal of scientific data analysis is to understand the behavior of a system or process based on a sample of the system. In many instances it is possible to observe both, input parameters and outputs, and characterize the system as a high-dimensional function. Such data sets arise, for instance, in understanding the uncertainty of large numerical simulations, energy landscapes in optimization problems, or the statistical analysis of image data relating to biological or medical parameters. In [GBPW10] we propose an approach that analyzes and visualizes such data sets. To do so it combines topological and statistical geometric techniques to provide interactive visualizations of discretely sampled high-dimensional scalar fields. The method relies on a segmentation of the parameter space using an approximate Morse-Smale complex on a cloud of point samples. For

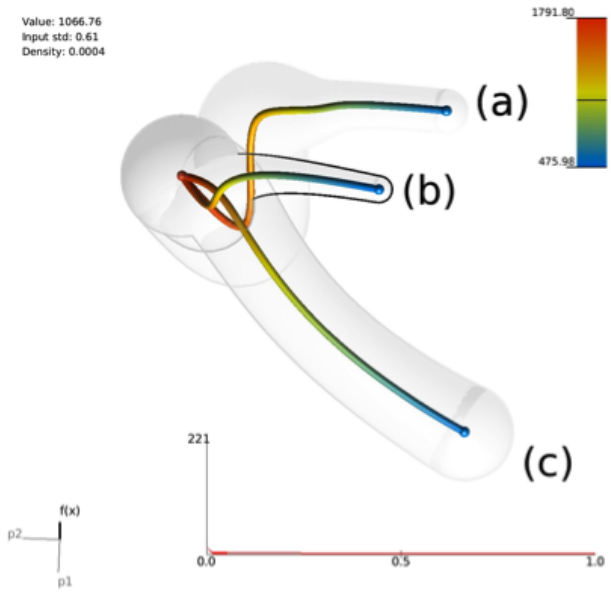


Figure 17. The three distinct minima (blue spheres) correspond to pure fuel, pure oxidizer and extinction/re-ignition. Graphs of chemical composition plotted against temperature for the crystals indicate that these three minima correspond to extinction (a), pure oxidizer (b) and pure fuel (c). The global maximum (red sphere) corresponds to efficient burning conditions.

each crystal of the Morse-Smale complex, a regression of the system parameters with respect to the output yields a curve in the parameter space. The result is a simplified geometric representation of the Morse-Smale complex in the high dimensional input domain. Finally, the geometric representation is embedded in 2D, using dimensionality reduction, to provide a visualization platform. The geometric properties of the regression curves enable the visualization of additional informa-

tion about each crystal such as local and global shape, width, length, and sampling densities. The method is demonstrated on several synthetic examples and several real scientific problems including: the analysis of manufacturing parameters and their effect on the strength of concrete, the parameters of climate simulations and their relationship to predicted global energy flux, the local concentrations of chemical species in a combustion simulation and their integrations with temperature, and the relationships between MRI brain images and measured clinical variables. Figure 17 shows our visualization technique applied to a high dimensional data set of chemical composition in relation to heat released during a jet flame combustion simulation.

4.2 Analysis of Large-Scale Scalar Data Using Hixels

One of the greatest challenges for todays visualization and analysis communities is the massive amounts of data generated from state of the art simulations. Traditionally, the increase in spatial resolution has driven most of the data explosion, but more recently ensembles of simulations with multiple results per data point and stochastic simulations storing individual probability distributions are increasingly common. In [TLB+11] we introduce a new data representation for scalar data called hixels that store a histogram of values for each sample point of a domain. The histograms may be created by spatial down-sampling, binning ensemble values, or polling values from a given distribution. In this manner, hixels form a compact yet information rich approximation of large scale data. In essence, hixels trade off data size and complexity for scalar-value "uncertainty". Based on this new representation we propose new feature detection algorithms using a combination of topological and statistical methods. In particular, we show how to approximate topological structures from hixel data, extract structures from multi-modal distributions, and render uncertain isosurfaces. In all three cases we demonstrate how using hixels compares to traditional techniques and provide new capabilities to recover prominent features that would otherwise be either infeasible to compute or ambiguous to infer.

Fuzzy Isosurfacing When down-sampling larger data sets, hixels enable the preserving the presence of an isosurface within the data. In particular, when hixels store the counts of all function values present within a block, we can use that to compute the likelihood of the presence of an isosurface within that block. Figure 18 demonstrates the results of this technique for a large combustion jet data set with half a billion grid points.

Sampling Topology Hixels encode the potential values along with their distributions at sample locations, a fact that can be exploited in visualizing the uncertainty in topological segmentations of down-sampled data. We use a sampling of the hixels to generate individual instances of the coarser representation, compute the Morse complex on the instance, and aggregate multiple instances of the segmentation to visualize its variability. We generate an instance V_i of the down-sampled data by picking values at each sample from the co-located hixel. The value is picked at random from the distribution encoded by the hixel. By picking values independently from neighboring values, we can simulate any possible down-sampling of the data, assuming all are independent.

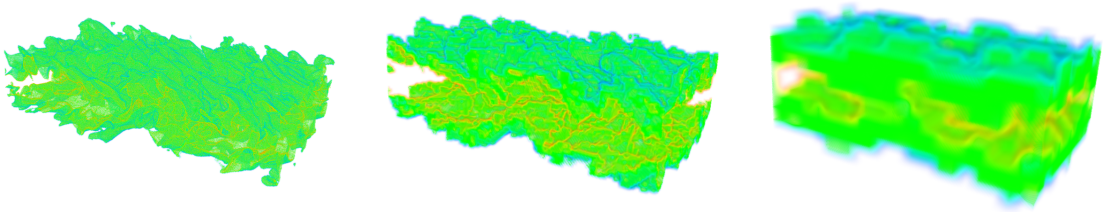


Figure 18. Volume rendering of the jet data set down-sampled using hixels. We visualize the scalar field g that indicates the likelihood of isovalue $\kappa = 0.506$ lying at that position. From left-to-right, we show hixels that block 2^3, 8^3, and 32^3 data values. Opacity is a triangle function centered at $g = 0$ and color is a rainbow map, red for high values, green for middle, and blue for low.

We perform convergence tests for a two-dimensional slice through a jet combustion simulation. In this experiment, we computed a hixel representation for the slice with blocks of size 8x8 and 16x16. We visualize in Figure 19 each aggregate slice for the 8x8 block size, as number of iterations and topological persistence are varied. The convergence of these sequences indicates that the distribution represented by the hixels produces implies stable modes of segmentation.

Extracting Structures from Multi-Modal Distributions As HPC resources increase, ensembles of runs are being computed more frequently to explore the state space of phenomena of interest. The resulting ensemble data comprises a collection of simulation results, each of which represents a state in the system defined by different input parameters and/or models. While ensemble data sets are hailed as a useful mechanism for characterizing the uncertainty in a system, their large size and variability pose significant challenges for existing analysis and visualization techniques. We developed a novel statistical technique for recovering prominent topological features from ensemble data stored in hixel format. This computation is aided by the fact that ensemble data has a statistical dependence between runs that allows us to build a structure representing a predictive link between neighboring hixels. Our algorithm identifies sub-regions of space and scalar values that are consistent with positive association and we perform topological segmentation on only those regions.

We demonstrate results on a mixture of 2 stochastic processes shown in Figure 20. This data highlights the fact that individual hixels can be multi-modal and can behave as both a minimum and maximum. A naive analysis that computes the mean or median of the hixels, followed by standard topological segmentation would fail to incorporate the multi-modal nature of the data. Our method addresses this issue by performing topological analysis directly on sheets of the domain that have likely simultaneously observable sets of behavior. Our approach clearly extracts separate sheets belonging to the two processes and identifies their prominent features.

To compare against down-sampling a large-scale data set, we also demonstrate results of this

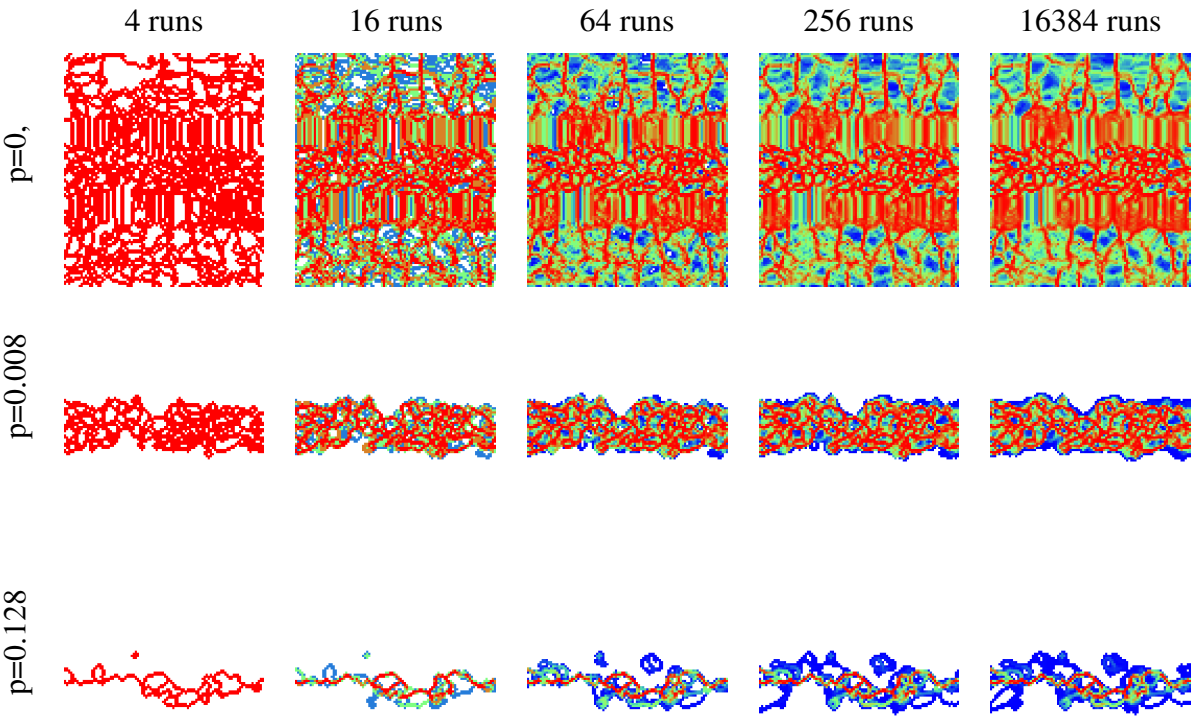


Figure 19. We sample the hixel data for an 8x8 blocking of combustion data, and compute the aggregate segmentation for a number of iterations, also varying the level of persistence simplification. Adjacent white pixels are identified in the interior of the same basin in every single run. The images converge as the number of iterations increases left to right.

method on a hixelated data set generated from the logarithm of the χ field of a lifted ethylene jet combustion data set with 1.3 billion grid points. The contingency tables between each pair of hixels are computed using observations between neighboring vertices along shared hixel faces. Figure 21 shows the number of buckets per hixel with block sizes of 16 (top-left), 32 (top-middle), and 64 (top-right). The color map ranges from blue at 1 bucket per hixel to red with 27 buckets per hixel and, as is to be expected, the number of buckets per hixel increases significantly as the block size increases. On the right the basins of maxima are shown for corresponding block sizes.

4.3 Feature-Based Statistical Analysis of Large Data

In [BKL+11] we present a new framework for feature-based statistical analysis of large-scale scientific data and demonstrate its effectiveness by analyzing features from Direct Numerical Simu-

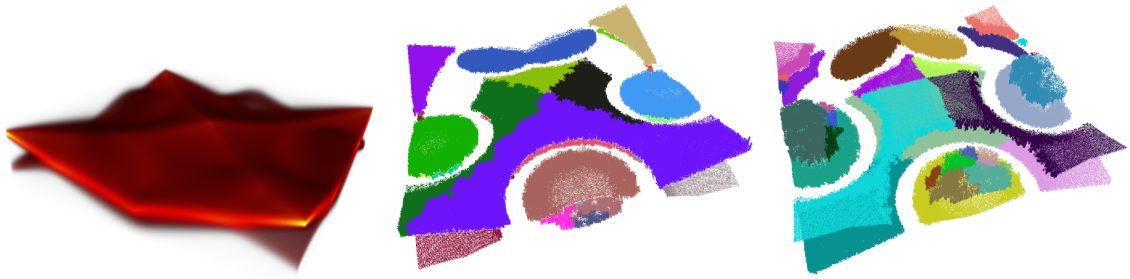


Figure 20. On the left, a volume rendering of a hixel data set generated by sampling images 3200 samples of a Poisson distribution and 9600 samples from a normal distribution.There are 512x512 hixels in this data set, each with 128 bins. The shortest axis in the images corresponds to histogram bins, thus a spatially higher location along that axis indicates a higher function value. Color and opacity are used to illustrate the density of samples. Thus the lower, right corner of shows a hixel with 2 distinct probable function values; the smaller function value is less probable than the larger. The center image shows basins of minima and the right image shows basins of maxima for this data set. By computing basins on sheets we are able to identify prominent features associated with each process in the mixture model.

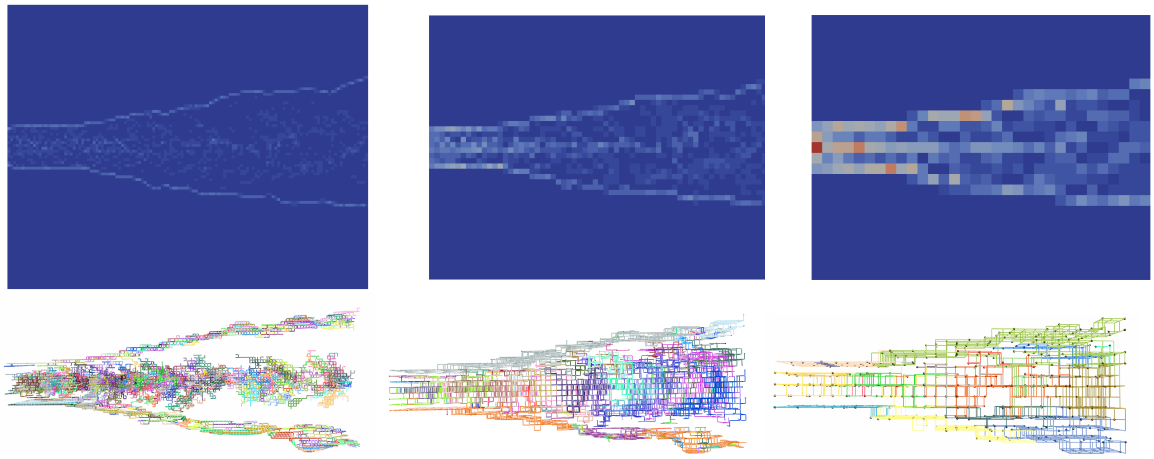


Figure 21. On the top the number of buckets per hixel is displayed for block sizes of 16 (left), 32 (middle) and 64 (right). Blue regions have 1 bucket per hixel while the maximum number of buckets per hixel is 27 and is shown in red. On the bottom the basins of maxima are shown for corresponding block sizes.

lations (DNS). Combustion scientists use DNS to study fundamental turbulence-chemistry interactions such as extinction and auto-ignition in turbulent jet flames. Of particular interest is the scalar dissipation rate, χ, which indicates the local rate of molecular mixing, which is enhanced by turbulent flow. Turbulent flow strains fluids, creating thin pancake-like features of locally high dissipation rate whose thickness provides a direct measure of the local mixing length-scale. Understanding the relationship between the thickness and the mean temperature within features is of principal interest to study the relationship between mechanical strains and chemical processes. This analysis is challenging due to the wide range of feature parameters that must be explored and the massive sizes of the simulation.

In our approach we pre-compute merge trees of the χ field which encode the set of features for all possible χ thresholds. Furthermore, we augment the merge trees with attributes, such as statistical moments of various scalar fields, e.g. χ, temperature, etc., as well as length scales computed via spectral analysis. The computation is performed in an efficient streaming manner in a pre-processing step and results in a collection of meta-data that is orders of magnitude smaller than the original simulation data. This meta-data is sufficient to support a fully flexible and interactive analysis of the features, allowing for arbitrary χ thresholds, providing per-feature statistics, and creating various global diagnostics such as cumulative density functions (CDFs), histograms, or time-series. We combine the analysis with a rendering of the features in a linked-view browser that allows scientists to interactively explore, visualize, and analyze the equivalent of one terabyte of simulation data on a commodity desktop. While we have successfully deployed our framework to analyze statistical properties of turbulent combustion, its design and implementation are general and applicable to a wide range of scientific domains.

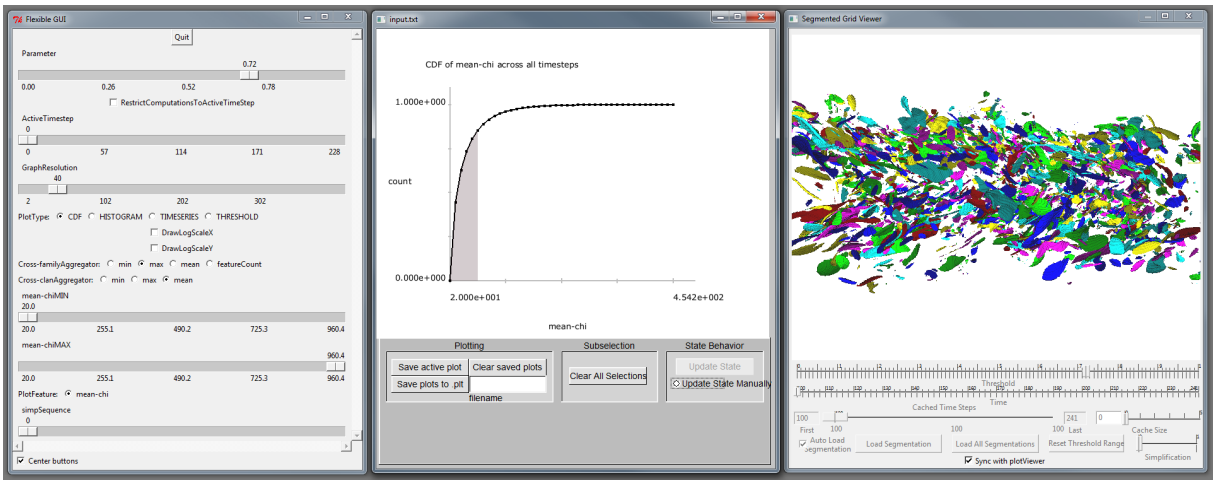


Figure 22. Our framework provides a natural and intuitive workflow for the exploration of global trends in feature-based statistics. By efficiently encoding hierarchical meta-data in a pre-processing step, interactive data exploration of the equivalent of one terabyte of simulation data is performed on a commodity desktop.

5 Dissemination of results

Refereed Publications

- **1 proceedings volume:** [GPRT11]
- **3 journal articles:** [GBPW10, PRSar, PRT11].
- **11 conference proceedings:** [nIRR10, BHPR11, RSST, BNPS10, WBGP, BJB+11a, GBP, GLP11, JBB+11a, PTB10, PTBM11, BBC+11]. (1 best paper, [BJB+11a])
- **1 multimedia submission:** [GLP11].
- **1 poster:** [BBC+11]

Submitted to Refereed Venues

- **8 journal articles:** [nIRRar, Hau11, Roj, RS10, BJB+11b, GKK+, BKL+11, JBB+11b]
- **1 conference proceedings:** [TLB+11]

Technical Reports

- [JBB+10]

Presentations

- **9 refereed conference presentations:** [nIRR10, BNPS10, WBGP, BJB+11a, GBP, GBPW10, JBB+11a, PTB10, PTBM11].
- **7 invited plenary presentations:**
  - Rojas: Banff International Research Station (Mar. 1, 2010)
  - Rojas: Notre Dame University (Aug. 4, 2010)
  - Pascucci: at the SIBGRAPI 2010 conference
  - Pascucci: IX Congress of the Peruvian Computing Society
  - Pascucci: Institute for Science and Technology (IST), Austria, January 27, 2011.
  - Pascucci: Commissariat à l'Énergie Atomique (CEA) TERA100 High Performance Computing Center, Arpajon, France , February 3, 2011.
  - Pascucci: Visualization in Computational Bioscience, February 24, 2011, Texas A&M University, TX.
- **10 invited conference/workshop presentations:**
  - Rojas: AMS regional meeting in Waco, TX (Oct. 18, 2009)
  - Rojas: (presented by postdoc Avendaño): International Symposium on Symbolic and Algebraic Computation, Munchen, July 28, 2010
  - Rojas: Toric Geometry and Applications, Leuven, Belgium (June. 6-11, 2011)
  - Pascucci: SCIDAC 2010 conference
  - Levine: IRTG meeting in Utah 2010
  - Bennett: BASC meeting in Palo Alto, CA, 2011

- Bennett: CScADS meeting in Lake Tahoe, CA, 2011
- Bennett: Grace Hopper Celebration of Women in Computing, Portland, Oregon, Nov. 2011
- Pascucci: 2nd National Conference in Advancing Tools and Solutions for Nuclear Material Detection, May 3, 2011, Salt Lake City, UT.
- Pascucci: Dagstuhl seminar on Scientific Visualization, Germany, May, 2011.

Sofware Packages (SAGE, MATLAB, C, C++)

- Scalable parallel statistics toolkit added to VTK, available at `git@vtk.org:VTK.git`.
- Demonstration of uncertainty in vector fields.
- Visualization of the structure of 3D Morse-Smale complexes.
- Computation of contingency statistics in arbitrary dimensions.
- Visualization of fuzzy isosurface in large and/or uncertain data sets.
- Identification of prominent topological features in uncertain data.
- Distributed computation of global statistics.
- Interactive visualization of discretely sampled high-dimensional scalar fields.
- Aggregation and visualization of feature-based statistics (length scales, descriptive statistics).
- Some preliminary code is available at `www.math.tamu.edu/~rojas/nearcircuits.html`

6 Planned Work

As we continue our work next year, we plan to investigate the following research activities:

- Build tools to support feature-based contingency statistics and demonstrate that they work at petascale, e.g., using large-scale combustion data.
- Build tools to identify modes in higher-dimensional distributions using topological techniques (begin with 2 and 3 dimensions, then extend to dimensions greater than 3).
- Turn select prototype implementations into reliable, scalable analysis tool kits for computational scientists to demonstrate generality of our approaches in an application-independent way.
- Compare hixels to existing compression techniques (e.g. wavelets).
- Apply hixel analysis methods developed this year to ensemble data sets.
- Fit and test non-Gaussian statistical models on topologically segmented features.
- Explore persistence as it pertains to hixelated data.
- Fix family of examples for experimentation with real solving.
- Complete Matlab and SAGE implementations of univariate chamber cone method.
- Implement a simple homotopy solver tailored to lower binomial systems and perform speed tests.
- Explore epsilon-cuttings for future speed-ups in the polyhedral portion of the chamber cone algorithm.

References

[BBC+11] Janine C. Bennett, Peer-Timo Bremer, Jacqueline Chen, Ray W. Grout, Andrea Gruber, Attila Gyulassy, Evatt Hawkes, Hemanth Kolla, and Valerio Pascucci. Conditional analysis of dns combustion data using local and global shape characteristics. In *2011 DOE Scientific Discovery through Advanced Computing (SciDAC) conference*, 2011.

[BHPR11] Osert Bastani, Chris Hillar, Dimitar Popov, and Maurice J. Rojas. Randomization, sums of squares, near circuits, and faster real root counting. *accepted for publication in an upcoming AMS Contemporary Mathematics volume*, 2011.

[BJB+11a] Harsh Bhatia, Shreeraj Jadhav, Peer-Timo Bremer, Guoning Chen, Joshua A. Levine, Luis Gustavo Nonato, and Valerio Pascucci. Edge maps: Representing flow with bounded error. In *4th Pacific Visualization Symposium*, pages 75–82, Hong Kong, China, March 2011. IEEE.

[BJB+11b] Harsh Bhatia, Shreeraj Jadhav, Peer-Timo Bremer, Guoning Chen, Joshua A. Levine, Luis Gustavo Nonato, and Valerio Pascucci. Flow visualization with quantified spatial and temporal errors using edge maps. *submitted*, 2011.

[BKL+11] Janine C. Bennett, Vaidyanathan Krishnamoorthy, Shusen Liu, Ray Grout, Evatt Hawkes, Jacqueline Chen, Jason Shepherd, Valerio Pascucci, and Peer-Timo Bremer. Feature-based statistical analysis of combustion simulation data. *conditionally accepted to IEEE Visualization 2011*, 2011.

[Ble06] Grigoriy Blekherman. There are significantly more nonnegative polynomials than sums of squares. *Israel J. of Math.*, 183:355–380, 2006.

[BNPS10] Matt Berger, Luis Gustavo Nonato, Valerio Pascucci, and Claudio T. Silva. Fiedler trees for multiscale surface analysis. In *IEEE INTERNATIONAL CONFERENCE ON SHAPE MODELING AND APPLICATIONS (SMI)*, 2010.

[BPRT09] J. Bennett, P. Pébay, D. Roe, and D. Thompson. Numerically stable, single-pass, parallel statistics algorithms. In *Proc. 2009 IEEE International Conference on Cluster Computing*, New Orleans, LA, August 2009.

[BPT09] J. Bennett, P. Pébay, and D. Thompson. Scalable k-means statistics with Titan. Technical Report SAND2009-7855, Sandia National Laboratories, November 2009.

[Cha01] Bernard Chazelle. *The Discrepancy Method*. Cambridge University Press, 2001.

[DCD86] D. Dacunha-Castelle and M. Duflo. *Probability and Statistics*, volume 1. Springer-Verlag, 1986.

[DFS07] Alicia Dickenstein, Eva Maria Feichtner, and Bernd Sturmfels. Tropical discriminants. *J. Amer. Math. Soc.*, 20:1111–1133, 2007.

[GBHP08] Attila Gyulassy, Peer-Timo Bremer, Bernd Hamann, and Valerio Pascucci. A practical approach to morse-smale complex computation: Scalability and generality. *IEEE Trans. Vis. Comput. Graph*, 14(6):1619–1626, 2008.

[GBP] Attila Gyulassy, Peer-Timo Bremer, and Valerio Pascucci. Towards topology-rich analysis and visualization. *Journal of Physics: Conference Series*.

[GBPW10] Samual Gerber, Peer-Timo Bremer, Valerio Pascucci, and Ross Whiteker. Visual exploration of high dimensional scalar functions. In *IEEE Transaction on Visualization and Computer Graphics*, volume 16, pages 1271–1280, 2010.

[GKK+] Attila Gyulassy, Natallia Kotava, Mark Kim, Charles Hansen, Hans Hagen, and Valerio Pascucci. Direct feature visualization using morse-smale complexes. *submitted*.

[GKZ94] Israel Moseyevitch Gel'fand, Misha M. Kapranov, and Andrei V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[GLP11] Attila Gyulassy, Joshua A. Levine, and Valerio Pascucci. Visualization of discrete gradient construction (multimedia submission). In *27th Symposium on Computational Geometry*, Paris, France, June 2011. ACM. Accepted.

[GPRT11] Leonid Gurvits, Philippe Pébay, Maurice J. Rojas, and David C Thompson. Papers from birs workshop on complexity, randomization, and relaxation. *Contemporary Mathematics*, 2011. In print.

[Har] Robin Hartshorne. *Algebraic Geometry*. Number 52 in Graduate Texts in Mathematics. Springer-Verlag.

[Hau11] Jon D. Hauenstein. Numerically computing real points on algebraic sets. *preprint*, 2011.

[Hen04] Amy Henderson. *The ParaView Guide*. Kitware, Inc., 2004.

[HS95] Birkett Huber and Bernd Sturmfels. A polyhedral method for solving sparse polynomial systems. *Math. Comp.*, 64(212):1541–1555, 1995.

[JB87] Carlos M. Jarque and Anil K. Bera. A test for normality of observations and regression residuals. *Revue Internationale de Statistique*, 55(2):163–172, 1987.

[JBB+10] Shreeraj Jadhav, Harsh Bhatia, Peer-Timo Bremer, Joshua A. Levine, Luis Gustavo Nonato, and Valerio Pascucci. Consistent approximation of local flow behavior for 2D vector fields using edge maps. Technical Report UUSCI-2010-004, SCI Institute, University of Utah, December 2010.

[JBB+11a] Shreeraj Jadhav, Harsh Bhatia, Peer-Timo Bremer, Joshua A. Levine, Luis Gustavo Nonato, and Valerio Pascucci. Consistent approximation of local flow behavior for 2D vector fields using edge maps. In *4th Workshop on Topology-Based Methods in Data Analysis and Visualization*, Zurich, Switzerland, April 2011.

[JBB+11b] Shreeraj Jadhav, Harsh Bhatia, Peer-Timo Bremer, Joshua A. Levine, and Valerio Pascucci. Quantized 2d vector fields. *submitted*, 2011.

[Kap91] Misha Kapranov. A characterization of $\mathscr{A}$-discriminantal hypersurfaces in terms of the logarithmic Gauss map. *Mathematische Annalen*, 290:277–285, 1991.

[Kit10] Inc. Kitware. *The VTK User's Guide, version 5.4*. Kitware, Inc., 2010.

[KOS09] K. Koizumi, N. Okamoto, and T. Seo. On Jarque-Bera tests for assessing multivariate normality. *J. of Statistics: Advances in Theory and Applications*, (1):207–220, 2009.

[LRW03] Tien-Yien Li, Maurice J. Rojas, and Xiaoshen Wang. Counting real connected components of trinomial curve intersections and m-nomial hypersurfaces. *Discrete and Computational Geometry*, 30(3):379–414, 2003.

[LV07] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer-Verlag, 2007.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. LeCam and J. Neyman, editors, *Proc. of the 5th Berkeley Symp. on Mathematics Statistics and Probability*, 1967.

[Mah36] P. C. Mahalanobis. On the generalised distance in statistics. *Proc. of the National Institute of Science of India*, (12):49–55, 1936.

[Nes03] Yuri Nesterenko. Linear forms in logarithms of rational numbers. *Diophantine approximation (Cetraro, 2000), Lecture Notes in Math.*, 1819:53–106, 2003.

[nIRR10] Martín Avenda no, Ashraf Ibrahim, Maurice J. Rojas, and Korben Rusek. Randomized np-completeness for p-adic rational roots of sparse polynomials in one variable. In *ISSAC 2010*. ACM Press, 2010.

[nIRRar] Martín Avenda no, Ashraf Ibrahim, Maurice J. Rojas, and Korben Rusek. Faster p-adic feasibility for certain multivariate sparse polynomials. In *Journal of Symbolic Computation, special issue in honor of 60th birthday of Joachim von zur Gathen*, To appear.

[PRSar] Mikael Passare, Maurice J. Rojas, and Boris Shapiro. New multiplier sequences via discriminant amoebae. *Moscow Mathematical Journal*, to appear.

[PRT11] Philippe Pébay, Maurice J. Rojas, and David C Thompson. Optimizing n-variate (n+k)-nomials for small k. *Theoretical Computer Science, Symbolic-Numeric Computation 2009 special issue*, 412(16):1457–1469, 2011.

[PT05] Mikael Passare and August Tsikh. Amoebas: their spines and their contours. *Idempotent mathematics and mathematical physics, Contemp. Math.*, 377:275–288, 2005.

[PTB10] P. Pébay, D. Thompson, and J. Bennett. Computing contingency statistics in parallel: Design trade-offs and limiting cases. In *2010 IEEE International Conference on Cluster Computing*, September 2010.

[PTBM11] Philippe Pébay, David Thompson, Janine Bennett, and Ajith Mascarenhas. Design and performance of a scalable, parallel statistics toolkit. In *12th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC-11)*, 2011.

[Roj] Maurice J. Rojas. Extremal sparse polynomial systems over local fields. *submitted for publication*.

[RS10] Maurice J. Rojas and Swaminathan Sethuraman. Refined asymptotics for multigraded sums of squares. *submitted to special issue of Theoretical Computer Science*, 2010.

[RSST] Korben A. Rusek, Frank Sottile, and Jeanette Shakalli-Tang. Dense fewnomials. *accepted for publication in an upcoming AMS Contemporary Mathematics volume*.

[Sto00] Arne Storjohann. *Algorithms for matrix canonical forms*. PhD thesis, Swiss Federal Institute of Technology, Zurich, 2000.

[TLB+11] David C. Thompson, Joshua Levine, Janine C. Bennett, Peer-Timo Bremer, Attila Gyulassy, Valerio Pascucci, and Philippe Pébay. Analysis of large-scale data using hixels. *submitted to the IEEE Symposium on Large-Scale Data Analysis and Visualization*, 2011.

[Ttn] Titan informatics toolkit. http://titan.sandia.gov/.

[WBGP] Gunther H. Weber, Peer-Timo Bremer, Attila Gyulassy, and Valerio Pascucci. Topology-based feature definition and analysis. In *Numerical Modeling of Space Plasma Flows: ASTRONUM-2010 ASP Conference Series*.

[Yat34] F. Yates. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

DISTRIBUTION:

2	MS 9018	Central Technical Files, 8944
1	MS 0899	Technical Library, 9536

Sandia National Laboratories