Comprehensible Presentation of Topological Information

Gunther H. Weber, Kenes Beketayev, Peer-Timo Bremer, Bernd Hamann,
Maciej Haranczyk, Mario Hlawitschka and Valerio Pascucci

Abstract

Topological information has proven very valuable in the analysis of scientific data. An important challenge that remains is presenting this highly abstract information in a way that it is comprehensible even if one does not have an in-depth background in topology. Furthermore, it is often desirable to combine the structural insight gained by topological analysis with complementary information, such as geometric information. We present an overview over methods that use metaphors to make topological information more accessible to non-expert users, and we demonstrate their applicability to a range of scientific data sets. With the increasingly complex output of exascale simulations, the importance of having effective means of providing a comprehensible, abstract overview over data will grow. The techniques that we present will serve as an important foundation for this purpose.

1 Introduction

Topology-based data analysis is applicable to a wide range of scientific data understanding problems, including mixing of fluids [10] and the analysis of combustion simulations [4]. The ability of topology-based method to provide a compact, abstract overview over complex data makes them an important tool for the analysis of exascale simulations. However, the information provided by topological structures is usually in the form of abstract graphs, and it is difficult for users without fundamental background in topology to interpret these graphs. Thus, its necessary to find new ways to present information so that it is more easily comprehensible without having an in-depth knowledge of topology. In this context, the use of metaphors, such as the toporrery [11], or topological landscapes [13] shows promising results. In this paper, we describe two ways to present topological information to users in an intuitive fashion. The first method is aimed at the analysis of optimization problems, where minima or maxima of a function are entities of main interest. Here, we focus on a problem from computational chemistry where the main interest focuses on transitions between states—characterized by minima. We have developed a method that uses the information encoded in the Morse complex to present a map-like representation of the system that highlights probable state transition [2]. While currently focused on chemistry, this type of representation can be useful for many optimization problems.

Another limitation of current presentation methods is that they focus on structural information. To alleviate this problem we have developed topological cacti. Topological cacti present the contour tree in an intuitive fashion, inspired by the toporrery representation. In addition to this structural perspective, topological cacti support displaying corresponding quantitative information. Both methods are useful tools in presenting topological information to an increasing audience of researchers interested in applying topology-based methods to their data and simplify the analysis of complex simulation results.

2 Map-based Representations for Analyzing Optimization Solution Spaces

Understanding the solution space plays an important role in a wide range of optimization applications. Using the topology-based analysis, it is possible to extract relevant information from high-dimensional cost functions. The Morse complex, e.g., encodes information about minima (or maxima) and relationships between them. We have developed a map-based [2] representation that shows this information in an intuitive fashion. While we utilized this representation in the context of computational chemistry, it has applications for a wide range of optimization problems.

Finding solutions to many urgent problems (e.g., pollution and global warning) requires novel insights into chemical systems and processes. It is possible to gain these insights through computational modeling and analysis of the system. One characteristic, common in analysis of many existing models, is energy as a function of coordinates of components of a chemical system. Analyzing this energy function (a specific instance of a cost function in optimization problems), defined on a high-dimensional domain, may improve our understanding of chemical systems and reactions.

There is a lack of effective visualization techniques for these multi-dimensional functions, and chemists often study only one or two parameters of interest at a time. Developing a comprehensive understanding of these complex systems requires new visualization techniques that show relationships between all parameters describing the system at the same time. The Morse complex contains the required information, but a background knowledge in topology is necessary to interpret it. To remedy this difficulty, we utilize a map-based metaphor that represents this information. Nodes in a graph represent minima (states) and edges transitions between those states. Edge width and color represent the likelihood of a transition. Maps include geometric/spatial information to show how distant solutions are by choosing appropriate node locations. Recasting the information contained in the Morse complex as states and transitions between them utilizes concepts familiar to chemists, and allows them to make effective use of it.

2.1 Extracting Topological and Geometric Information

Structural information encoded in the Morse complex. To obtain structural information describing transitions, we approximate the Morse complex for the energy function f_e by producing a segmentation of the input data into stable/descending manifolds. We implemented this segmentation similar to the approach of Gerber et al. [9] by finding a neighbor with the steepest descending gradient for each point, and using a union-find data structure to determine its corresponding minimum. Each stable state of the chemical system corresponds to a minimum in the energy function. The Morse complex is a partition into regions such that following the gradient from any point in the region will end at the minimum, i.e., such that segmented regions correspond to classes of configurations associated with a stable state.

Each descending region can adjoin several other regions. The saddle between two adjoining regions, i.e., the lowest function value along their boundaries, corresponds to the barrier between the states. Since minima and saddles in the Morse complex correspond to energy minima and barriers, chemical transformation paths are given by the edges connecting two minima through a saddle. Assigning minima and saddles to nodes and connecting pathways as edges creates the Morse complex graph. Based on this relationship between stable states and barriers, we construct a graph that represents the system.

Simplifying the structural perspective. Noise in a dataset along with increased data complexity can result in convoluted final graphs, and we need effective means of simplifying the Morse complex graph before analyzing it. We have implemented a scheme that eliminates low-persistence minima, where persistence is the absolute difference of energy function values. Since one minimum might belong to several minimum-saddle pairs, and thus have multiple associated persistence values, we select the lowest among them to decide whether a minimum is eliminated.

We consider two approaches for finding a suitable persistence threshold. The conventional approach is based on a persistence diagram obtained by gradually increasing the simplification threshold value within the possible value range and determining the current number of minima for each value. Big drops in the number of minima indicate a large number of related minima and thus candidates for appropriate simplification threshold values. While this approach can be very useful due to its generality, we also utilize an alternative calculation based on chemical domain knowledge about the considered system.

Preserving distance information. Whereas the graph itself is an abstract structure, we have to find a proper layout to visualize it. Although standard graph-layout algorithms may produce visually pleasing visualizations of the Morse complex graph, this representation does not consider relational information of the original system, which is important for chemical analysis due to impact of distances on the probability of certain transformations.

To preserve this relational context, we project the high-dimensional point locations of the graph onto the plane. Depending on the type of input data and the required results, there exist a variety of methods to reduce the data dimensionality [8]. Considering the fact that our points are provided in an Euclidean space and we are using a linear scaling function, the method simplifies to what is commonly known as principal component analysis (PCA). Using the principal components of the spread of points the new coordinate system ensures a good preservation of distances, which is important for understanding the data.
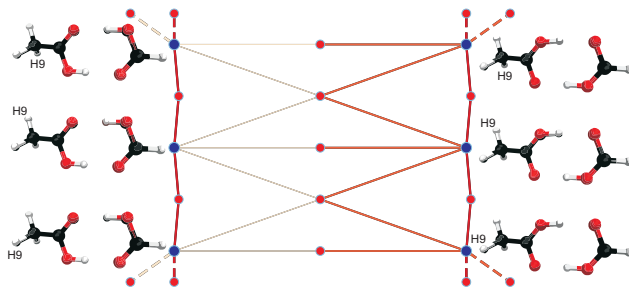
Figure 1: A 2D graph showing the topology of the 3D potential energy function of a complex of formic and acetic acids, which depends on the positions of constituting atoms. The strength of this approach is that it shows an entire high-dimensional energy landscape in terms of stable states and transitions between them, which are concepts familiar to chemists.

2.2 Visual Representation as Map

While it seems intuitive to apply multi-dimensional scaling both to minimum and saddle positions, we found that this approach often leads to a cluttered graph layout. Projecting only minima and placing saddles along lines between minima results in a much cleaner graph layout and supports encoding additional information based on the location of barriers. While the projection at this point is simple to navigate, overlaps of edges and nodes might conceal important information. Thus, we implemented an iterative optimization algorithm that minimizes node/edge overlap in the final layout.

Once we have obtained final positions for all nodes of the graph, we add chemical information from the model system. Since the most important information is the energy difference between a minimum and a barrier, we color edges according to the energy difference value. The Boltzmann distribution states that the probability of accessing a state decreases exponentially with its energy. The lower the energy difference is, the higher is the probability of a transformation. Hence, we display paths with smaller energy differences more prominently (darker colors and wider edges). We note that a higher energy difference leads to very narrow and lightly colored edges, making them almost invisible. These edges can be neglected when drawing the final layout, and we further simplify the graph by discarding edges above a threshold signifying the highest possible and/or interesting energy difference. Periodic edges, i.e., edges crossing a boundary along which the problem is periodic, are dashed, and the common saddle of the two connected minima is broken into two nodes (see Figure 1). Finally, we add labels to all minima that specify coordinates (first line) and energy value (second line).

2.3 Application Examples

Figure 1 shows a graph of the topology of the 3D potential energy function of a complex of formic and acetic acids, which depends on the positions of constituting atoms. Each sample in this energy function corresponds to a simulation run for a particular configuration of the system. Blue and red dots represent minimum energy configurations and lowest barriers connecting neighboring minima, respectively. Edges represent the energy cost of a particular transformation, with darker and wider edges corresponding to transformations through lower barriers (more likely transformations). Two vertical branches corresponding to different positions of protons are visible on the left and right side of the graph. Energy barriers for transforming the right branch into left one are lower than for the reverse transformation.

The free energy function of CH_4 in an LTA zeolite is substantially different from the energy functions of the dimer of formic acid and acetic acid, see Figure 2ab. This graph represents only the large cage of LTA zeolite, which is the only fragment of void space in LTA accessible to CH_4. The graph representation of the free energy reveals important information about the material. There are 14 important energy minima per periodic unit cell of LTA corresponding to favorable locations of the adsorbing CH_4 molecule (adsorption sites). Six of them correspond to lower energy (ca. 3.7 k_BT), and are localized near windows connecting two periodic cells (near faces of the unit cell). The remaining eight minima are localized further away from the windows, on the surface of the large cage of LTA. The further analysis of
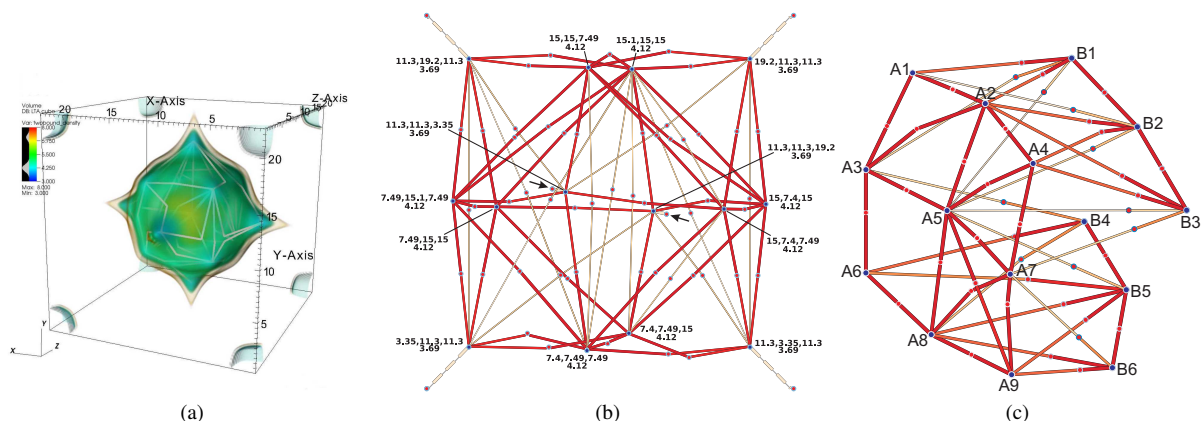
Figure 2: (a) Volume rendering of the energy function of a CH_4 molecule in an LTA zeolite and lowest energy paths connecting neighboring minima. (b) Corresponding graph showing lowest energy paths (an edge going through a face of the periodic box marked with arrows). (c) Graph of the 4D energy function of DFA. Nodes are labeled A1 to A9 and B1 to B6 to highlight two subgraphs. A1-A9 correspond to minima with coordinates of 0.95Å along the first and second dimension. The remaining third and fourth coordinates are (45,345), (180,345), (60,180), (285,345), (180,180), (60,0), (300,180), (180,0), and (300,0) for A1 to A9, respectively. Similarly, the first and second coordinate for B1-B6 are 1.75Å. The remaining coordinates are (60,345), (180,345), (300,345), (60,0), (180,0), (300,0) for B1 to B6, respectively.

connections between nodes/minima in our representations reveals that all 14 minima localized within the big cage are separated by low barriers, and therefore hops of CH_4 between the adsorption sites are feasible. However, connections between large cages in the extended material lead through high barriers. These high barriers along diffusion paths in every direction are reflected in slower diffusion rates.

The final example is the graph representation of the 4D energy function of DFA depicted in Figure 2c, for which plotting the original energy functions in all four dimensions was not feasible. The resulting graph is fairly easy to analyze. It clearly shows energy separation of the graph into two subgraphs: one corresponding to rotamers of DFA (with minima denoted A1-A9) and another with double proton transfer (minima denoted B1-B6). Transitions between minima within each subgraph are at low cost as shown by heavy edges.

2.4   Impact and Outlook

Our approach holds promise in allowing chemists to gain insight into complex chemical systems and understand higher-dimensional energy functions. Since topological features of the Morse Complex correspond directly to concepts such as stable states familiar to chemists, it provides an intuitive overview of a chemical system. In the long term, it will allow chemists to analyze higher-dimensional systems at once without having to look at pairs of parameters separately (see Figure 2c for a four-dimensional example). We also plan to adapt this representation to other optimization problems, such as providing an overview over the solution space in autotuning of codes for high-performance architectures.

3   Combining Structural and Quantitative Information

Current visualization techniques provide either structural or quantitative information based on topological analysis over a data set. Contours, the connected components of level sets, play an important role in understanding the global structure of a scalar field. In particular their nesting behavior and topology—often represented in form of a contour tree–have been used extensively for visualization and analysis. Traditional contour tree visualizations only encode structural properties like number of contours or the nesting of contours, but no quantitative information such as volume

or other statistics. To gain full understanding of data sets, it is desirable to combine these complementary views in a single visualization. We introduced a new visual metaphor for contour trees, called topological cacti [14], that extends the traditional toporrery display of a contour tree to display additional quantitative information as width of the cactus trunk and length of its spikes. To demonstrate the effectiveness of this approach, we have applied the new technique to scalar fields of varying dimension and different measures.

3.1 Background and Related Work

Contour trees. The contour tree [3] is a special case of the more generally defined Reeb graph [12] and captures the topological evolution of an isosurface as the isovalue varies. Critical points, where the number of contours changes, appear as nodes in the contour tree. Nodes of degree one (leaves of the tree) correspond to extrema, where contours are created or destroyed. Interior nodes of degree three or higher correspond to "saddles," where two or more contours merge or split. Arcs of the contour tree represent contours between critical points, i.e., contours that do not change topology (with the exception of genus changes) as the isovalue varies between critical values.

Contour tree simplification. Contour trees of even moderately-sized data sets can be quite complex due to noise or the presence of features at various scales. To remove noise, or features at smaller scale, it is necessary to simplify the contour tree [7, 11]. Simplification ranks features—connected isosurface components—based on geometric measures, such as persistence (difference in function value) or volume signifying the "importance" of a feature, and removes those features that are deemed unimportant in this order. The branch decomposition [11] is an efficient data structure for encoding a multi-resolution representation of the contour tree. It decomposes the contour tree into a set of branches, where a root branch connects a minimum-maximum pair in the contour tree. Other extrema are ordered by an importance metric, and child branches connect them to a saddle in an existing branch. This representation makes it simple to access the contour tree at a desired level of detail.

3.2 Computing Per-Component Measures

Contour spectrum-based measures. The contour spectrum [1] computes global measures for an isosurface as a function of isovalue. Based on the relationship between a tetrahedral mesh and simplex splines, the authors derive a B-Spline representation of isosurface area as a function of isovalue within an individual tetrahedron of the data set. Combining the B-Spline functions of all individual tetrahedra of a given mesh results in a B-Spline representation of isosurface area for the entire data set. Integrating this area b-spline over isovalue yields B-Spline describing the volume enclosed in the isosurface.

To display per-component information about an isosurface, we associate it with a previously computed contour tree [6] represented as branch decomposition [11]. In the original contour spectrum algorithm, contributions of all tetrahedra are accumulated into a global B-Spline. To compute area on a per-component basis, we associate a B-Spline (or alternatively a piecewise polynomial function) with each branch of the branch decomposition, and accumulate contributions per-component. Linear tetrahedra do not contain any critical points—particularly saddles—in their interior. Consequently, a tetrahedron corresponds to a single path between two vertices (not necessarily nodes) in the contour tree, along which the function value uniquely identifies the contour containing this tetrahedron. As a result, we can consider each tetrahedron and accumulate its contributions to the branch decomposition.

Computing per-contour volume is more difficult. If a contour contains another contour, it is ambiguous, whether the volume of the nested contour should be subtracted or not. To avoid ambiguities, we follow the convention that we calculate volume from the last saddle. Thus, the volume associated with a contour corresponds to the volume swept by that contour as long as it "maintained its identity," i.e. since its last "interaction" with another contour. Adding the contribution to the branch decomposition is similar to adding the contribution for the area. There are two key differences: Instead of stopping at the maximum value of the tetrahedron, the contribution on the maximum branch reaches until the next saddle value since once the maximum value of the tetrahedra is passed the volume function is a constant corresponding to the volume of the tetrahedron. Furthermore, following our convention, for each segment, we subtract the volume at the saddle from the entire function, such that for each segment only the volume swept since the saddle is considered.
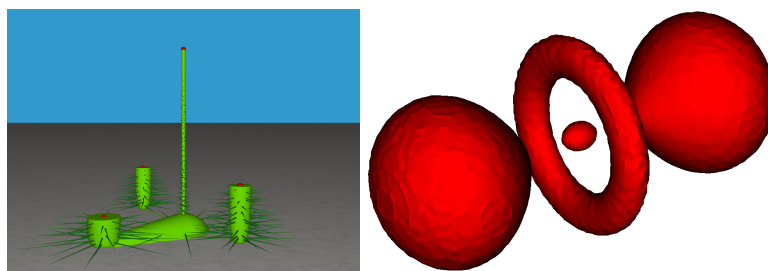
Figure 3: Topological cactus for the hydrogen atom (left) and isosurface rendering of the corresponding data set (right). The root branch in the center corresponds to the contour in the center of the ring. The two symmetric branches correspond to the two lobes, and the lowest branch to the ring. The branch width corresponds to the cube root of enclosed volume, and spike length to the square root of surface area.

Segmentation-based measures. In many application areas, it is desirable to display additional statistics, such as the mean value of a second variable on the isosurface. We compute these values at discrete locations of the contour tree or the join tree. To compute these values, we consider the segmentation implied by the contour tree [15] or the join tree [4] and accumulate these statistics for all samples corresponding to a branch or arc. This calculation would yield a single value per branch. To approximate function behavior along a branch or arc, we split it at regular intervals and insert degree two nodes into the join tree.

3.3   Topological Cacti Layout

We use a radial graph layout identical to that of the toporrery [11] for generating topological cacti. Extrema are ordered in concentric discs, where the disc is chosen based on the depth of an extremum in the branch decomposition. Like in the original toporrery, the radius ratio is constant in our layout, and each branch is assigned a radial wedge with an angle proportional to the number of child branches. To display metrics, we draw each branch as cylindrical extrusion and chose the radii according to a first selected measure (clamping however at a minimum radius to ensure that branches do not disappear). We draw connectors as sphere sweeps to ensure smooth connection of the cylindrical extrusion corresponding to parent and child branch. Before mapping a metric such as volume to radius, we support choosing a variety of transformation functions (square root, cube root, logarithm) to ensure that higher-dimensional properties are appropriately scaled to the one dimensional radius.

For display of a second metric, we augment the cylindrical extrusions with spikes. We start by placing circles along regular intervals of the cylindrical extrusion and choosing angles in regular intervals along this circle. We then choose spike location by jittering these regular samples, i.e., we add random offset angles and heights to the originally chosen spike locations. Spikes are drawn as cones with a fixed (user defined) base radius and a length depending on the metrics. Similar to the extrusion radius, we support user chosen mappings such as square root, cube root or logarithm.

3.4   Application Examples

Figure 3 shows an example of a topological cactus for the hydrogen atom data set (simplified to a persistence of ten). This is a simple, small—128^3 byte samples—example data set that we use to illustrate the concept. We display volume (mapped via cube root) as branch width and surface area (square root) as spike length, to examine how whether persistent features also correspond to large features. The root branch corresponds to the evolution of the contour in the center. It is immediately obvious that despite its very high persistence (resulting in it being chosen as root branch), both its volume and surface area are very small. The two lobes, corresponding to the two symmetric branches are much larger in both volume and surface area. Finally, the ring, corresponding to the lowest branch, has the largest volume and area.

Figure 4a shows the topological cactus of the merge tree of fuel consumption rate in a combustion simulation (one 4.3G time step from a coarse AMR simulation). Here, we are interested how the size of burning regions correlates with
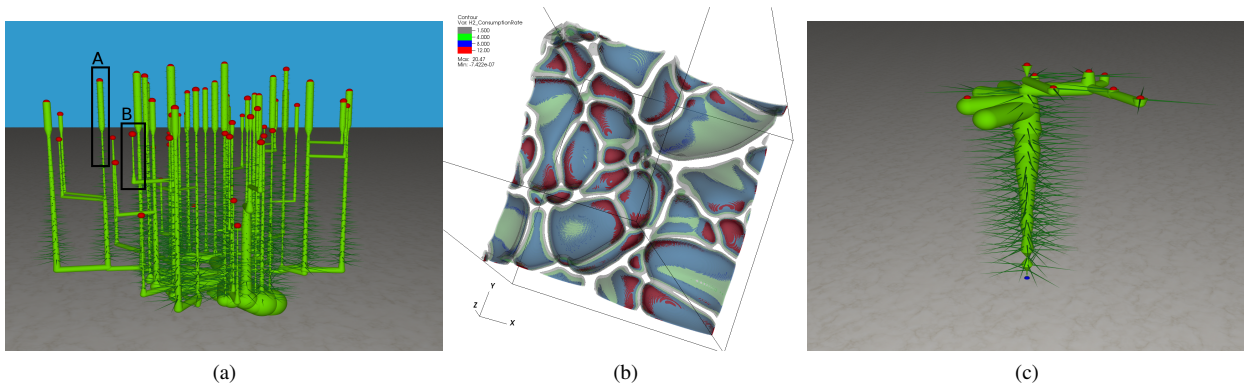
Figure 4: (a) Topological cactus of the merge tree of fuel consumption rate in a combustion simulation (a). The width corresponds to the cube root of number of vertices in a region. Spike length shows the variance of temperature. (b) Isosurfaces for four different fuel consumption rates illustrating burning regions for these thresholds. (c) Topological cactus of a climate simulation showing the merge tree of long wave energy output in the winter months (left). The width corresponds to cube root of vertex count. The spike length shows the mean value of FLUT.

temperature variance. Thus, we map the cube root of volume—approximated as the number of vertices—to branch width and temperature variance to spike length.

The cactus shows that there are two types of regions: (i) Branches that are relatively wide close to the maximum (rectangle "A" in the figure marks an example) and thus correspond to regions comprised of a relatively large number of vertices and (ii) branches that are thin close to the maximum (rectangle "B" in the figure marks an example) corresponding to small regions. The latter occur at lower fuel-consumption levels (i.e., their maxima are generally lower than those of the wide branches). The spikes show that temperature variance is small where there is high fuel consumption, which occurs where there is "intense burning" (i.e., where fuel consumption is high). Furthermore, the temperature variance is (expectedly) larger for non-burning regions (i.e., regions of no to low fuel consumption).

Figure 4c shows applying our method in the context of uncertainty quantification. The cactus show the merge tree of a 21 dimensional climate simulation ensemble. The data set consists of 1197 climate simulations performed with the Community Atmospheric Model (CAM) [5]—each using a different combination of 21 input parameters. Such ensembles are used to study the sensitivity of the climate predictions to changes in input parameters. The output function shown in Fig. 4c is the average longwave energy output (FLUT) in the winter months (December, January, February). The cactus shows that most of the "volume" of the parameter space is collected in several large but relatively low persistent branches of maxima. This fact is interesting as the observed energy output is expected to be closer to the global minimum. Thus, it appears that only a small portion of the input samples produces realistic outputs. Nevertheless, all persistences are fairy low indicating that the overall behavior is stable under perturbation of the input parameters.

3.5   Impact and Outlook

Topological cacti combine insights obtained from the contour tree with metrics derived on a per contour bases. This combination can greatly enhance the usefulness of contour tree representations, in particular when importance measured by persistence differs from other metrics such as area of volume. The contour spectrum implementation that associates contour spectrum with contour tree has also applications in other topological data analysis settings.

Going forward, we will improve the contour spectrum calculation. Our current approach of computing the contour spectrum in piecewise polynomial/B-Spline representation works well for quantized data sets. However, for real-valued data, each data value appears in the knot vector, making this representation inefficient. To counteract this problem, we need to devise a simplification scheme that can be applied during the calculation and still support error bounds.

Acknowledgements

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

References

[1] Chandrajit Bajaj, Valerio Pascucci, and Daniel R. Schikore. The contour spectrum. In *Proc. IEEE Visualization*, pages 167–175, 1997.

[2] Kenes Beketayev, Gunther H. Weber, Maciej Haranczyk, Peer-Timo Bremer, Mario Hlawitschka, and Bernd Hamann. Visualization of topology of transformation pathways in complex chemical systems. *Computer Graphics Forum (Special Issue, Proceedings Eurographics/IEEE Symposium on Visualization)*, pages 663–672, May/June 2011. LBNL-5019E.

[3] Roger L. Boyell and H. Ruston. Hybrid techniques for real-time radar simulation. In *Proc. 1963 Fall Joint Computer Conference*, pages 445–458. IEEE, 1963.

[4] Peer-Timo Bremer, Gunther H. Weber, Julien Tierny, Valerio Pascucci, Marcus S. Day, and John B. Bell. Interactive exploration and analysis of large scale turbulent combustion using topology-based data segmentation. *IEEE Transactions on Visualization and Computer Graphics*, 17(9):1307–1324, September 2011. LBNL report number pending.

[5] Community atmosphere model (CAM). See http://www.ccsm.ucar.edu/models/atm-cam/.

[6] Hamish Carr, Jack Snoeyink, and Ulrike Axen. Computing contour trees in all dimensions. *Computational Geometry—Theory and Applications*, 24(2):75–94, 2003.

[7] Hamish Carr, Jack Snoeyink, and Michiel van de Panne. Simplifying flexible isosurfaces using local geometric measures. In *Proc. IEEE Visualization 2004*, pages 497–504, 2004.

[8] Imola Fodor. A survey of dimension reduction techniques. Technical report, LLNL, 2002.

[9] Samuel Gerber, Peer-Timo Bremer, Valerio Pascucci, and Ross Whitaker. Visual exploration of high dimensional scalar functions. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), November 2010.

[10] Daniel Laney, Peer-Timo Bremer, Ajith Macarenhas, Paul Miller, and Valerio Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE Transcations on Visualization and Computer Graphics*, 12(6):1053–1060, 2006.

[11] Valerio Pascucci, Kree Cole-McLaughlin, and Georgio Scorzelli. *The Toporrery: Computation and Presentation of Multi-Resolution Topology*, pages 19–40. Springer-Verlag, 2009.

[12] G. Reeb. Sur les points singuliers d'une forme de pfaff complètement intégrable ou d'une fonction numérique. *Comptes Rendus de l'Acadèmie des Sciences de Paris*, 222:847–849, 1946.

[13] Gunther H. Weber, Peer-Timo Bremer, and Valerio Pascucci. Topological landscapes: A terrain metaphor for scientific data. *IEEE Trans. Vis. Comp. Graph. (Proc. Visualization 2007)*, 13(6):1416–1423, 2007.

[14] Gunther H. Weber, Peer-Timo Bremer, and Valerio Pascucci. *Topological Cacti: Visualizing Contour-based Statistics*, pages 63–76. Springer Verlag, Heidelberg, Germany, 2011. LBNL-5018E.

[15] Gunther H. Weber, Scott E. Dillard, Hamish Carr, Valerio Pascucci, and Bernd Hamann. Topology-controlled volume rendering. *IEEE Trans. Vis. Comp. Graph.*, 13(2):330–341, 2007.