ASCR Exascale Research Kick-Off Meeting—Project Handout

Project: Topology-based Visualization and Analysis of Multi-dimensional Data and Time-varying Data at the Extreme Scale

PI: Gunther H. Weber; Computational Research Division and National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory; Computer Science Department, UC Davis

Project Members: Dmitriy Morozov (Postdoctoral Researcher)

Problem Description

Computing at the extreme scale makes it possible to simulate physical phenomena of unprecedented complexity that comprise a growing number of dependent model variables. The results of these simulations are already beginning to exceed our capabilities for analyzing them effectively. Without aggressive improvements in data analysis technology, we will not be able to analyze effectively future simulation results and derive new insights from those simulations. In particular, there is a need to extract features that are not explicitly present in the data, such as burning regions in combustion simulations or storm systems in climate applications, and provide researchers with quantitative information about these features.

Methods to be Employed

We will extend topology-based data analysis to tomorrow's data understanding problems. Topology-based methods have proven effective on today's problems, e.g., in identifying and analyzing features in various science domains including combustion and porous materials, and they offer hope for meeting data understanding challenges as data grow larger. Topology provides qualitative geometric information by studying the general concept of "connectedness." Like clustering-based techniques, this information can help to identify data categories, e.g., in high-dimensional data sets. However, in a majority of data-understanding problems that we will consider, features are closely related to isosurfaces. This relationship makes topology-based methods a prime candidate for helping to define and extract high-level features. Examples include burning regions in combustion simulations or "fingers" of high carbon concentration in carbon sequestration simulations, which usually can be characterized as connected regions. Topological information is also insensitive to distance metrics and coordinate system choices, which is important in high-dimensional data understanding problems where no natural choice for metrics or coordinate systems exists. Furthermore, the construction of "summaries" over whole parameter space domains is at the basis of many topological techniques, and topological methods can help in identifying appropriate parameters. Finally, topological structures can serve to abstract data and reduce the amount of data that needs to be processed, saving compute time during analysis. We will concentrate our work on three focus areas (i) adapting current (2D and 3D) topology-based methods to massively parallel architectures, (ii) using topology-based methods for *in-situ* data analysis and (iii) applying topology-based methods to high-dimensional data sets to demonstrate their applicability and appropriateness for this use case.

Topological Methods at the Extreme Scale We will develop novel algorithms that implement current (3D and 2D) topology-based data analysis methods effectively on massively parallel architectures. We will also collaborate with researchers performing large-scale simulations on large DOE computation platforms and use simulation data produced by them to test our analysis algorithms. The main goals for this research area are (i) applying topological analysis to the result of at least one heroic run on a large DOE computation

platform, (ii) identifying potential bottlenecks in the scaling of topology-based data analysis algorithms and (iii) providing a practical means of applying topology-based data analysis to results of hero runs.

***In-situ* Topological Data Analysis** Many simulations already produce more data than it is feasible to write to disk, and this trend will likely be exacerbated by future supercomputing architectures. As a consequence, it may not be possible to write all simulation data to disk for subsequent analysis, which decreases data analysis accuracy. We will investigate the use of *in-situ* topological analysis to improve data analysis accuracy. We will further utilize topology-based methods to identify relevant and interesting behavior in a simulation and use this information to control what simulation results are written to disk. Our goal is to utilize this integration to write additional, topology-based information to disk that will improve analysis accuracy. We will also investigate file-formats that use topological information to reduce the amount of data written to disk while still supporting later effective analysis and utilizing topology-based methods to control which times-steps are written to disk. The main goals in this research area are (i) a prototype of implementing topological data analysis within the BoxLib/AmrLib framework; (ii) using the new framework to analyze combustion simulations; (iii) demonstrate that this framework supports effective analysis on a greatly reduced amount of data written to disk; (iv) utilize topological analysis as part of the simulation to control what time steps are written to disk.

Multidimensional and Multivariate Feature Mining using Topological Methods With the increasing complexity of simulations, it becomes increasingly important to analyze multivariate and/or multidimensional simulation results. We will demonstrate that topological structures like contour tree, Reeb graph and/or Morse complex can be utilized to gain insight in high-dimensional data. Furthermore, we will devise simplification schemes that are required to eliminate noise and eliminate features at small scale that are not of interest. We will also derive quantitative information about detected features (such as size distribution) and devise visualization methods that, e.g., present the graph structure of the contour tree and Reeb graph in a way that facilitates comparison of simulations. The main goals for this research area are (i) demonstrating applicability of topological data analysis to high-dimensional data for one or more application areas; (ii) determining for what application areas topological analysis of high-dimensional data proves particularly useful and (iii) identifying commonalities between these areas, particularly with respect to definition of features of interest, derived quantities of interest and effective presentation means.

Impact

The proposed research will help to bridge the growing gap between advances in simulation complexity and current data analysis capabilities by providing researchers with powerful topology-based data analysis tools that run on tomorrow's parallel machines. These tools will help them in deriving new knowledge from simulations and significantly improve our ability to analyse simulation results. A recent report on a DOE Workshop on "Mathematics for Analysis of Petascale Data", for example, recognized that need and listed "the challenges of high dimensional data" as a key theme for DOE relevant research, and our proposed research will address this urgent need. Implementing topological data analysis *in-situ* will improve the fidelity of data analysis for time-varying simulations where it is not possible to write all data to disk. Furthermore, it will provide means to adapt the amount of data written to disk based on the presence of features-of-interest, thus maximizing the portion of data relevant to analysis in the simulation output.