Exploring Uncertainty Visualization in Large Data Sets (Renewal Proposal)

Alex T. Pang
Computer Science Department,
University of California, Santa Cruz

Executive Summary: This is a proposal to renew and extend the current project with LANL on uncertainty visualization of large data sets. The size of the data sets and the uncertainty in the data sets come from the fact that we are dealing with ensemble data sets. These are usually from monte carlo simulations where each output (out of many runs) represents a possible solution. The degree of agreement (or disagreement) provides some indications of certainty (or uncertainty) about the results. Because monte carlo simulations can potentially involve large number of repetitions, the total data size can very quickly get very large even for relatively small spatial and temporal resolutions.

During the first year, we focused on gridded ensemble data sets. Based on discussions from a LANL visit in November 2007, they are more interested in studying ensemble data sets from particle based simulations. These ensemble data sets do not have an explicit grid structure, but rather, comes as a set of "point clouds". While the underlying visualization strategy remains, handling these type of data sets require special data structure and preprocessing for efficiency. Another feedback from the November visit is the need to tailor our research and development to support specific applications rather than a general methodology for dealing with ensemble data sets. Thus, we plan to address these two needs during the second year: (i) extend methodology to support point cloud ensemble data sets, and (ii) demonstrate applicability of methodology by focusing on a specific application and data set to be determined LANL and follow-on discussions. For the latter, one possibility would be working with data from cosmological simulations that James Ahrens had previously worked on in the context of comparative simulation. If that's the case, the scientist partner would then be Katrin Heitmann (http://t8web.lanl.gov/people/heitmann/). In conjunction and anticipation of this, I have also contacted and got the support to work on Stan Woosley's data set. Stan is leading a 5 university, 3 national lab consortium, funded through DOE's SciDAC program, to study supernovae and gamma ray bursts. He also has ongoing collaborations with LANL.

Project Description: The motivation for this project remains the same – the national laboratories run large scale numerical simulations in support of scientific inquiries into different fields such as high energy physics, physical processes in the earth's mantle, dynamics of mixing fluids, etc. Several enabling technologies are required to provide support for this mode of scientific investigation including parallel and supercomputing, petascale data storage and management, as well as visualization and analysis tools to help scientists understand their data.

This proposal will focus on visualization tools and techniques. In particular, we will focus on a particular form uncertainty representation and their subsequent visualization.

Dealing with and accounting for uncertainty is an important component in scientific experiments. Uncertainty can be introduced in the form of simplified models, insufficient numerical precision and stability, instrumentation drifts and miscalibration, propagation of errors in simulations, etc. Uncertainty comes in various forms and have multiple facets. The following examples can be considered one form of data uncertainty or another: missing data, conflicting data, sparse data, data with poor pedigree, data from repeated measurements, data from numerical experiments with simplified physics, coarse resolution, different initial and boundary conditions, etc. Previously, and even to this date, the most common way of representing uncertainty in data is either with a single scalar value representing concepts such as data quality, reliability, standard deviation, etc. or with a pair of numbers representing the data range.

The drawback with either approach is the inability to discriminate and probe the uncertainty further. For example, standard deviation assumes the data comes from a normal distribution and can fail otherwise i.e. a bi-normal, or a multi-modal distribution data can have the same standard deviation as a normally distributed data set, and yet the underlying uncertainty will be drastically different. Even if the data did come from a normal distribution, the standard deviation alone is insufficient in characterizing the probability of uncertainty.

To address this obvious deficiency, we propose to use the entire probability density function (pdf) to represent both the data and its associated uncertainty. These pdf's on different physical variables can be built from repeated measurements, but in the national labs context, more typically from multiple numerical simulations e.g. with monte carlo simulations. For example, an 3D time varying ocean dynamics simulation may involve tens of variables including salinity, pressure, temperature, current, vorticity, etc. Each simulation run would therefore produce a multi-dimensional (3 spatial + 1 temporal), multi-variate (however many physical variables are of interest) data set. Now, do monte carlo simulations with populations of a few hundred runs, and the data size quickly gets quite large. Furthermore, current visualization techniques typically ignore the richness of such data sets, and usually resign to using summary statistics such as standard deviation.

The challenge of this proposal is to create visualization techniques for analyzing such spatio-temporal pdf data sets. That is, at every point in space-time, there is a pdf for each physical variable. We refer to such data type as a *multi-value*. So, a 3D temperature pdf volume, would result in a 3D scalar multi-value; while a 3D current pdf volume, would result in a 3D vector multi-value. Our proposed approach for analyzing multi-value data sets is with an operator approach where an operator can either be mathematically or procedurally defined. Basic operations such as addition and multiplication of multi-values need to be defined, as well as comparison operators for multi-values. These would form the basic building blocks for rudimentary tasks such as interpolating 2 multi-values, or comparing 2 multi-values. Note that operators are broadly defined and can take advantage of domain specific techniques e.g. if the multi-value is treated as a time-series rather than as a pdf of a random variable, signal processing formulations could be used an operators. As noted in the summary, based on the feedback from our LANL visit, we need to apply operators that are more application relevant. One such possibility that we will investigate is applying the idea of Logarithm of Odds [1] to multi-value data set. An attractive property is that the LogOdds have mathematical closure for addition and multiplication. For example, the probabilistic addition of two discrete distributions p_1 and p_2 is carried out in LogOdds space by adding LogOdds(p_1) and LogOdds(p_2) and then mapping the result back using the inverse LogOdds^{-1} operation.

Aside from investigating application specific and relevant operators to use with our multi-value visualization methodology, we also plan to extend the approach to work with point cloud ensemble data sets. Unlike gridded data sets, point clouds do not have implicit neighborhood information. Conceptually, every point in the simulation run will have information about its current spatial location, time stamp, as well as values for all the relevant physical variables. The number of particles and their locations may be different from one ensemble member to another. Even when dealing with a single ensemble member, visualizing large point clouds can be quite challenging. This can be further compounded when the variables of interest are vectors instead of scalar values where data access patterns needed to generate appropriate visualizations do not solely rely on spatial proximity. There are several approaches that one can use to support such data sets. The list usually includes space partitioning, hierarchical grouping, and resampling to a grid. Some of these techniques may be better suited for scalar data fields that are primarily static. We plan to investigate these methods in the context of visualizing ensemble point cloud data sets where we are interested in the velocity field. A possible direction is to see if there are other partitioning strategy that may be better suited for our target application. For example, if there's some structure in the spatial organization of the point cloud, one could create a spatial partitioning based on a single instance of an ensemble point cloud, then apply the partitioning to the entire ensemble. This can be justified because ensemble members are usually statistically very similar to each other. Another possibility might be some physics-based partitioning where groupings are made according to the physical behavior of the system e.g. in tracking particles in an expanding spiral galaxy, it may be more advantageous to form groups along the spiral arms rather than splitting groups into sector wedges. On the other hand, splitting neighboring particles around a black hole, which acts as a sink, into wedges may be more advantageous. In this project, we plan to investigate approaches that would facilitate tracking the movement of particles (probabilistic streamlines) that can take advantage of intelligent partitioning and thereby improved pre-fetching and caching of needed data.

Project Status: During the first year, our research focus was on a visualization strategy for probabilistic streamline tracing of ensemble data sets. Towards this end, there were 2 questions that we seeked to answer:

1. How does one generate the "streamline" from a given initial seed point? One possible avenue is to use importance sampling at each integration step and generate a probabilistic "cone" trajectory. Regions within such a cone can be delineated with different probability of occurrence as well.
2. Given a seed point and another point in the flow field, what is the probability that a particle flows from one to the other point? We're still not sure what would be a physically reasonable approach yet. One possibility is to take the union of both a forward and backward integration pass.

Most of our results are related to question 1. Figure 1 shows some of our results so far. The second question is actually a very important problem in the domain of fiber tracking in DT-MRI (diffusion tensor magnetic resonance imaging). The fiber tracks represent white matter fiber bundles in the brain. The fiber connectivity (where it starts and ends) have important significance to functional brain mapping. I have recruited the help of another graduate student – Yeon Gwack to help in this regard. We have obtained a high angular resolution diffusion image (HARDI) data set from a brain researcher from UCSF. Typically, a DT-MRI data set is obtained from 7 diffusion images (each one is a 3D scalar volume) which are used to calculate a 3D 2nd order tensor representation – hence the term diffusion tensor. The data set that we obtained from UCSF contains 133 diffusion images. The problem with the 2nd order tensor representation is that there's ambiguity in tracing the fiber when the track approaches an isotropic (no directional preference) part of the data. We plan to use the HARDI data set to generate multiple 2nd order tensor representations, and apply our probabilistic streamline tracing approach on the ensemble of 2nd order tensor fields. As far as we can tell, this approach has not been tried before in the medical imaging and visualization communities.

The premiere venue for reporting visualization research is the IEEE Visualization conference. The paper deadline for this year's conference, to be held in Ohio, is March 31, 2008. We are shooting for a paper that will report on the results from this project at this venue.

Personnel: The graduate student funded on this project is Eddy Chandra. Eddy is pursuing his MS degree with James Davis, and is planning to complete his degree requirement around June 2009. I recruited him to work on a visualization project from Summer 2007 as he did not have funding at that time.

During the summer of 2007, Eddy got up to speed on several aspects of the research problem:

1. Ability to work with a multi-dimensional, multivariate, multi-valued ocean circulation data set from Harvard (we were not able to get a sanitized data set from LANL to work with).

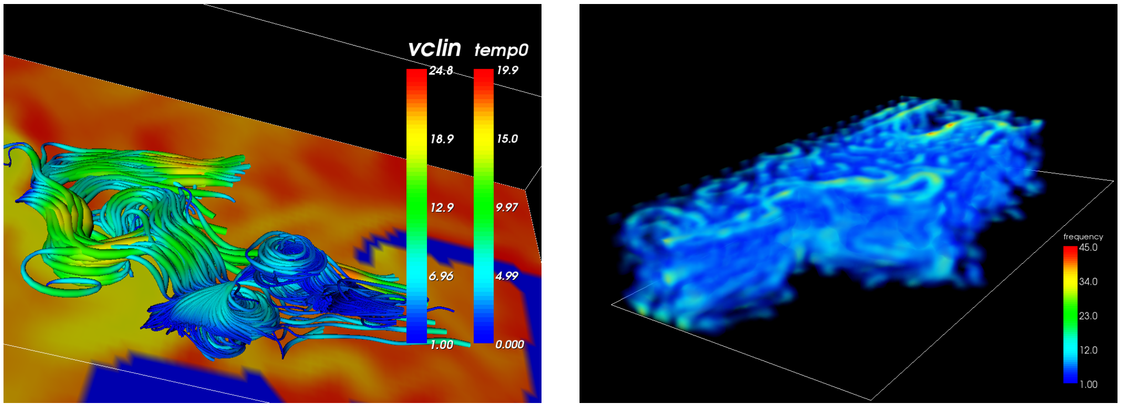

Figure 1: *Left*: Streamlines from an ensemble vector field showing overall similarity of flow patterns, yet subtle differences from different ensemble members. *Right*: Volume rendering showing the frequency of streamlines crossing a volume element. Regions with coherent structures showing where streamlines seem to congregate are readily visible even though the algorithm did not explicitly look for these.

2. Got up to speed on standard techniques for flow visualization using both Vtk and his own implementation – the latter is needed since the off-the-shelf technique is insufficient to handle the multi-value data set.

3. Also got up to speed and implemented a number of standard statistical techniques such as histogram, density estimates, calculation of cdf, and importance sampling of a pdf.

4. Finally, he also implemented a baseline, brute force method of looking at the multi-value vector field (essentially, a ”spaghetti plot” of the vector field).

This quarter, Eddy made more progress, and I started to give him additional reading material on ”probabilistic streamline tracing” (from the field of DT-MRI field). The most significant achievement he made this quarter is the implementation of a “volume frequency” and its subsequent direct volume rendering. The volume frequency is simply a grid superimposed on top of the data grid that is used to count the number of streamlines passing through a voxel. The number of streamlines going through a voxel is then used as an indication of “likelihood” that streamlines from an ensemble data will go through that particular voxel. Eddy’s current implementation uses individual streamline calculations from a small sample of the ensemble population. He has not calculated the volume frequency from using probabilistic streamline calculations directly yet. But this is the next thing on his agenda. The reason why I think this is the most significant achievement so far is because the resulting images show some coherent structure in the data set that we had not anticipated before. We are also waiting to hear back from the scientist (Pierre Lermusiaux) from Harvard on whether the structures are scientifically significant or not. In the meantime, Eddy is double checking his code to make sure that everything is in order.

Overall, I am fairly happy with Eddy’s progress. Given that I had to look for a student on short notice and he was willing to step up to the plate even though it’s not his main research interest. I was a bit hesitant on giving him some of the more technical reading materials (on DT-MRI) earlier since he had quite a bit of catching up to do, plus the material is quite heavy in the math (tensor manipulations). Hopefully, with the initial success and experience, Eddy will be more excited about this line of work. I would recommend that we continue to fund him.

On a related note, during my visit to Los Alamos, we had quite a few scientists (other than CS types) who attended my presentation. I was told by my contact (James Ahrens) that one of them was fairly high (2nd or 3rd level manager). The general indication is that the level of interest on visualizing ensemble data sets and uncertainty is fairly high. On the other hand, the ensemble data sets of the scientists that were there also indicated that their simulations are usually particle based and do not have explicit grid information. Therefore, my plan is once Eddy completes the work for this year – main goal right now is for an IEEE Visualization paper submission in March 2008, he could focus on extending the methodology and techniques to work on particle based ensemble data sets. Eddy also indicated that he's amenable to spending summer 2008 at LANL.

Future Plans: We have submitted an NSF proposal in response to their peta-scale computing program jointly with colleagues from UC Davis. We are planning to submit another joint proposal to an upcoming DOE call for proposal which is suppose to come out "any time now".

Lab contacts: The primary LANL contact is Jim Ahrens, visualization team leader of the Advanced Computing Laboratory (ACL). A possible new collaborator on the project may be Katrin Heitmann.

1 References

[1] Kilian M. Pohl, John Fisher, Sylvain Bouix, Martha Shenton, Robert W. McCarley, W. Eric Grimson, Ron Kikinis, and William M. Wells. Using the logarithm of odds to define a vector space on probabilistic atlases. *Medical Image Analysis*, 11:465–477, 2007.