Toward Visual Analysis of Ensemble Data Sets

Andrew T. Wilson
Sandia National Laboratories
P.O. Box 5800, MS 1323
Albuquerque, New Mexico 87185-1323
atwilso@sandia.gov

Kristin C. Potter
SCI Institute, University of Utah
Street Address Here
Salt Lake City, Utah 12345-6789
kpotter@cs.utah.edu

ABSTRACT

The rapid and continuing increase in available high-performance computing resources has driven simulation-based science in two directions. First, the simulations themselves are growing more complex, whether in the fidelity of the models, spatiotemporal resolution or (more frequently) both. Second, multiple instances of a simulation can be run to sample the results of parameters within a given space instead of at a single point. We name the results of such a family of runs an *ensemble data set*. In this paper we discuss the properties of ensemble data sets, consider their implications for analysis and visualization algorithms, and present a few insights into promising avenues of investigation.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Scientific Databases*; I.3.8 [**Computer Graphics**]: Applications; G.3 [**Probability and Statistics**]: Statistical Computing

General Terms

Delphi theory

Keywords

ensemble data sets, visualization, statistics, scientific data management, parallel databases

1. INTRODUCTION

Simulations are playing an ever-increasing role in the study of physical phenomena because of the advantages they offer over laboratory experiments. A simulation is:

- **Repeatable**: Instead of preparing new experimental samples one need only re-run the software.
- **Free of physical constraints**: Any environment may be modeled no matter how large, small, hostile or physically plausible.
- **Measurable**: Data values can be extracted at every simulation point instead of at a tiny handful of sensor positions.
- **Not tied to real time**: Fast phenomena may be slowed down and slow phenomena may be accelerated. Events that happen on nanosecond scales are as accessible as those that take millions of years to evolve.

At the same time, however, the nature of a simulation necessarily brings disadvantages. A simulation is further:

- **Inaccurate**: Its results can only be as accurate as the underlying physics equations and numerical approximations.
- **Coarse**: The necessarily finite resolution always omits detail that may have profound effects on the results.
- **Fallible**: In addition to the approximations in the physics implementation, bugs in the simulation code are very difficult to rule out completely.

Nonetheless, for phenomena that are difficult or impossible to produce in the laboratory, simulations are often the best available tool to explore the consequences of a given set of starting conditions under a given model. They come fully into their own when we conduct "what if" experiments to guide processes of design and decision-making. In this paper we discuss the challenges of analyzing ensemble data sets and offer our opinions on some of the primary research issues. We illustrate the discussion using examples from recent investigation into climate and weather analysis tools [10].

1.1 Predictive Capability and Uncertainty

One increasingly common and important use case for simulations is the exploration of phenomena with uncertain input conditions. This uncertainty can be *aleatory*, or fundamentally irreducible, arising from quantities that cannot be known in advance; or *epistemic*, arising from a lack of data that could (potentially) be remedied with more measurements [3]. In either case the goal is to predict a set of possible outcomes in spite of the uncertainty. Given a set of samples from the space of possible parameters and the simulation output corresponding to each set of inputs, we can consider the aggregate of the results to be an approximation of the space of possible outcomes. We refer to such aggregates as *ensemble data sets*.

1.2 Organization

In section 2 we discuss in more detail what ensemble data sets are and what distinguishes them from any other simulation result. In section 3 we outline common driving questions that lead to the

Ultrascale Visualization Workshop 2009, Portland, Oregon USA

creation of ensemble data and the research issues that those questions imply. In section 4 we offer a few insights gleaned from early work with weather and climate ensembles. We conclude in section 5.

2. WHAT IS AN ENSEMBLE?

In this section we go into more detail about the properties of ensemble data sets in order to frame the driving questions and research issues discussed in the rest of the paper.

Ensemble data sets typically arise as part of a predictive analysis of some real-world phenomenon with uncertain inputs. The process begins with the real-world physical processes being modeled and input parameters such as material properties, physical constants and ambient conditions. Without full documentation for all of physics, every step beyond this starting point is necessarily an approximation that introduces error. The result is one or more numerical models of the process of interest and one or more sets of input parameters (called *input decks* for convenience). An ensemble data set is created when each model is executed using each input deck. We refer to the individual simulation results that compose the ensemble as its *members*. The result is used to gain insight into the likely outcomes of the simulation given the uncertainties in the input.

An ensemble data set typically has all of the following properties.

- Ensembles are **large**. Small data sets such as the NOAA/NCEP Short-Range Ensemble Forecast [1] yield about 20GB of data each time they are executed. Larger systems such as global climate models can easily generate hundreds of terabytes of results. These numbers continue to grow without bound.

- Ensembles are **multivariate**. The SREF ensemble contains more than 400 state variables sampled at every grid point. Global climate simulations often have on the order of 100 state variables and tens of input parameters. Thermal or mechanical finite-element simulations often have fewer than 10 parameters and outputs.

- Ensembles are **multivalued**. By their nature they contain multiple values for each variable at each point – one per member of the ensemble. Given enough information about the input uncertainty, these multiple values can be considered as samples of a probability distribution function (PDF) at each output point. There is as yet little work in visualization for PDF-valued data. Most efforts [6, 8, 4] have focused on the meaning of the set of values in the context of the analytical task instead of applying a general approach to a specific task. We consider this a strong argument for domain specificity as we will discuss later.

- Ensembles are **expensive to generate and store**. Because they incorporate multiple runs of a simulation, they multiply both the computation time and the storage requirements of a single simulation by the number of members in the ensemble.

- Ensembles are **time-varying**. All of the common use cases that lead to ensemble data involve the evolution of a system over time. Moreover, the transient events in the system are often of greater interest than the end state or instantaneous maximum of some variable.

- Ensembles are **awkward**. There are few deployed tools with the capability to handle the whole of an ensemble with reasonable interactivity and analytic power. Scientists will often pick a few locations that they deem good surrogates for the behavior of the whole, extract data from those locations, and discard the rest of the ensemble. The resulting files are trivially small and easy to manipulate but neglect the wealth of information and context that was thrown away. This can be dangerous if the choice of surrogate points was incorrect.

2.1 How Ensembles Mitigate Uncertainty

Uncertainty arises in every phase of the simulation process from the initial model through to the (hopefully rare) data errors caused by undetected hardware errors. In such situations any single simulation has a fundamental limitation: it gives results only for a single numerical model and a single set of parameters and initial conditions. However, we can use ensemble data sets to bound and mitigate the effects of such uncertainty in the following ways.

- **Multiple models**: There are usually many choices for how to design a simulation. For example, finite element models can use Lagrangian or Eulerian meshes. Climate models can incorporate different equations for land, sea, atmosphere and ice. Weather simulations may incorporate multiple models to balance out the strengths and weaknesses of any one. There may be a computationally inexpensive model that is accurate enough for most situations and a far more detailed one that can cover the difficult cases. An ensemble data set can accommodate any or all of these cases to combine the strengths of different approaches.

- **Multiple grids**: In any simulation there is often a spatial or temporal resolution that is "good enough"; that is, increased resolution will not yield increased detail. Ensemble data can reduce the uncertainty due to insufficient resolution by demonstrating convergence to such a sufficient resolution.

- **Multiple inputs**: Using different input decks for each member of an ensemble serves two functions. First, by sampling the breadth of an input parameter space, the outputs can convey the breadth of possible responses as well as the likelihood of each one. However, it is generally not possible to enumerate the entire output space. This is troublesome in cases where high-impact events occur in small, low-probability, unvisited regions of the input space. Second, in simulations that start from measured data, the inputs can sometimes be deliberately skewed to compensate for weaknesses in the numerical models.

3. RESEARCH FRAMEWORK

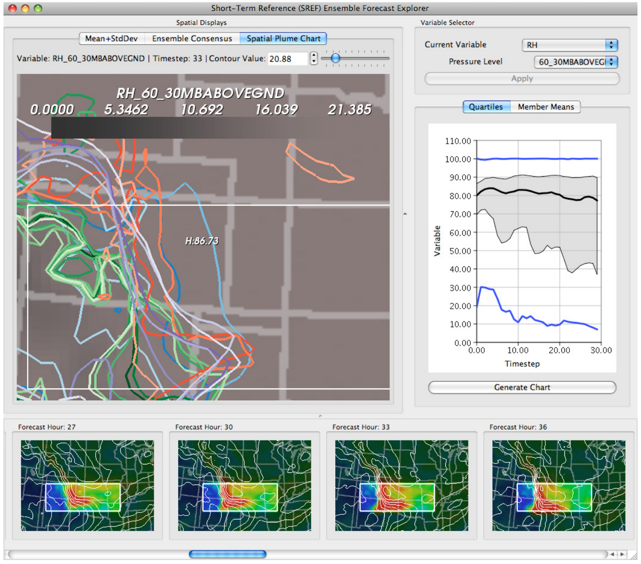
Since ensembles are used mainly as a tool to reduce and mitigate the effects of uncertainty, we can observe common factors in the typical analytical questions that lead to their creation. This leads in turn to a set of common research issues that occur across different analytical domains. In this section we outline common elements in analytical questions across domains and discuss broadly the research issues that arise as we begin to address them.

3.1 Driving Questions

The most basic question common to all ensemble analysis tasks is the one posed by the simulation itself:

What conditions or events are predicted by this range of possible input conditions?

The vagueness of the question suggests that we cannot address it directly at this level of abstraction. Instead, concrete answers must



Variable: RH_60_30MBABOVEGND | Timestep: 33 | Contour Value: 20.88

Current Variable: RH

Pressure Level: 60_30MBABOVEG

Figure 1: Prototype weather ensemble analysis tool incorporating multiple views. By displaying multiple facets of the data at once – summary statistics at right, multiple timesteps at bottom and comparative isocontours at top left – we convey a more complete picture than any one display metaphor could alone.

arise from within each concrete domain as in the following examples. In climate simulations we may be interested in surface temperature and sea level. In finite-element simulations of car crashes we may look for the maximum stress on critical components. In weather simulations during summer we may care most about severe storms.

Although this overarching question is too abstract to be tractable without domain knowledge, some of the questions that support it immediately suggest algorithms. Moreover, those algorithms resemble database queries. For example:

Where and when does some condition occur in the ensemble? We can answer this by finding node or element IDs where all the specified conditions occur in the state variables. This is a simple Boolean query evaluated node by node during a scan through the ensemble. In many cases it can even be answered using summary statistics or by the kind of range index constructed by FastBit.

What is the relative probability of some set of conditions? An answer to this question subsumes an answer to the above. The relative probability at any point in the ensemble can be computed as the number of member data sets that satisfy the analyst's conditions divided by the total number of member data sets. However, this will usually require a full scan through the data since the summary statistics are necessarily too coarse.

What conditions lead to some event of interest? Whereas the previous two questions can be satisfied with simple scans through the data, this one requires exploration. The event itself can presumably be detected with the methods described above. However, the conditions leading to it are usually a higher-order phenomenon inferred from the data instead of being directly present in the ensemble. For example, in an analysis of projected rising sea levels, we might choose to look first at which ice masses melted to contribute to the rise. Given that answer we could drill down further to ask what caused temperatures in those regions specifically to rise. The chain of reasoning continues until the analyst has as complete an explanation as desired or possible.

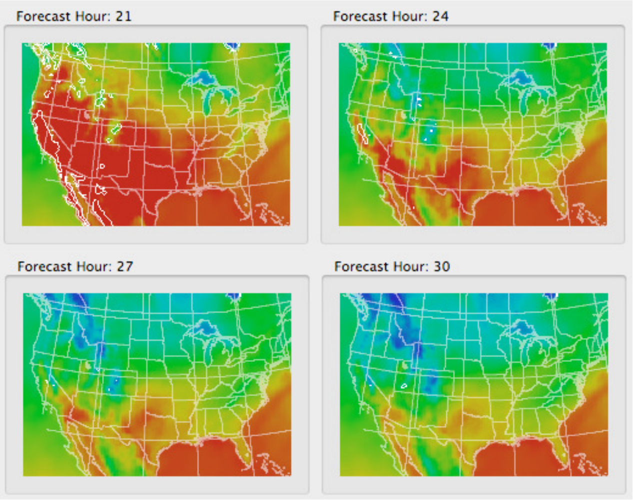


Figure 2: By displaying multiple timesteps side-by-side instead of in sequence we make it easier to simultaneously observe changes in several different areas of the data. This image shows cooling across the western United States as simulated time crosses from day to night.

What events occur that the analyst did not expect? Discovering the unexpected is a purely exploratory task. Since it depends wholly on the analyst's beliefs regarding the ensemble, it is incumbent upon the analyst to search through the data visually and numerically to verify that those beliefs are correct – or at least not contradicted.

3.2 Research Issues

The driving questions above suggest a set of capabilities necessary for scalable analysis and visualization of ensemble data sets. We treat these in no particular order. Our opinions on the most effective ways to treat each issue are set out in section 4.

3.2.1 Data Management

The sheer size of ensemble data sets poses major challenges. Specifically, the 'simple linear scan' we refer to in the previous section can take many hours – a cost incompatible with the notion of interactive analysis.

3.2.2 Many-Valued Data

When considering the data values themselves we can view an ensemble from two perspectives. On one hand, each state variable can have a collection of time-varying scalar (or vector or tensor) values at each point. On the other hand we can claim that each state variable has only a single value at each point and that the value is a distribution. The former view is most useful when we want to examine the behavior of individual ensemble members. The latter comes into play when we consider aggregate behavior.

3.2.3 Multidimensional Data

An ensemble is properly a multidimensional data set. It comprises (usually) two or three spatial dimensions, one dimension of time and one-to-many of the input parameter space. However, We only have two dimensions available for display. We must somehow reduce and project the data until a 2-dimensional representation can capture and exhibit a property of interest.

4. A FEW INSIGHTS

We maintain the assumption that the actual data for the ensemble is too large to fit into main memory. Moreover, since it is not possible to render all the data on screen, let alone inspect it directly, doing so is inefficient. Our approach in working with ensemble data has been to precompute summary statistics such as the moments about the mean and approximate quantiles [13, 7] that will fit in main memory. We then use those statistics to generate as much of what the user sees as possible. In many cases such as the view shown in Figure 3 the summary statistics are entirely sufficient to present an accurate view to the user. When queries arise that cannot be answered accurately with just the summary statistics, we rely on indexing schemes (including those summary statistics) to restrict I/O to only those parts of the data that we actually need.

In this section we discuss lessons we have learned from building tools for ensemble analysis for finite-element, weather, and climate data.

4.1 Storage and Retrieval

We think of the data store containing an ensemble as being more like a relational database with indices and query capability than as a serialized representation of simulation data structures. We are willing to trade a certain amount of storage overhead for efficient random access, especially if these storage schemes allow multiresolution or stream-like access to the raw data as in Pascucci and Frank [9]. We can pursue the database metaphor further with indexing schemes such as FastBit [12] that accelerate range queries.

The danger of summary structures and indices is that as they grow more detailed they can become large enough to pose data management problems in their own right. We believe that the most effective way to address these problems for the original ensemble data as well as acceleration or summary structures is by bringing hardware parallelism to bear. Here we consider immediately available approaches as well as ongoing research.

First, if we take the database metaphor literally then we can apply off-the-shelf parallel database warehouse hardware and software. These systems partition a data set across many disks and structure query execution to minimize the amount of data movement between disks. Database warehouse appliances were designed for business analytics queries whose hallmark is relatively simple computation over very large volumes of data with complex schemata. While this model does not match ensemble analysis perfectly we have found it expressive enough to efficiently execute all the queries we have encountered so far. Another advantage of this approach is that there are several companies that design and market parallel database appliances. As of late 2009 these include Netezza, TeraData, XtremeData and Greenplum, among others.

Second, we can use frameworks such as MapReduce [2] and Hadoop [11] to scatter the processing load across clusters of commodity hardware. This is most suitable in situations where we already have a cluster with enough local disk capacity to hold the entire ensemble. Like the relational model, MapReduce is not a perfect fit for ensemble analysis but is expressive enough to handle nearly all of our queries.

Finally, we anticipate the results of research efforts aimed specifically at creating repositories and tools for scientific data management such as the SDM Research Center at Lawrence Berkeley Laboratory. While these efforts are not as immediately available as the existing software and hardware mentioned above, in the longer run they will provide us with tools well adapted for the particular needs of scalable ensemble analysis.

4.2 Data Manipulation for Visualization

4.2.1 Distributions as Point Values

Visualization of spatially- and time-varying PDF-valued data is still an area of open research [6, 8, 4, 5] Moreover, extant work in this area concentrates on relatively small data sets from the perspective of ultrascale visualization. It is unclear how well existing algorithms will scale to data sets with millions or billions of elements. We believe that like other multiresolution algorithms we will need some sort of aggregate encompassing a set of underlying distributions. A simple sum is attractive but will also suppress small, interesting features. The literature on topology- and curvature-preserving geometric simplification provides helpful inspiration here.

4.2.2 Multiple Dimensions

As stated before, we must somehow reduce an ensemble data set to two dimensions before it can be displayed on a screen. We think of this process as the repeated application of one of three operators, each of which eliminates one or more dimensions at a time. Those operators are as follows:

Select: Eliminate a dimension by choosing a single value (e.g. a single timestep or ensemble member) and extracting a slice through the ensemble at that value.

Aggregate: Eliminate a dimension by ignoring it. For example, when computing a mean at a given point we ignore the ensemble member ID and average across all the members. Similarly, when computing summary statistics for a region we ignore both the ensemble member ID and spatial location.

Project: Eliminate one or more dimensions by projecting into a lower-dimensional subspace using methods like singular value decomposition (SVD). While this is more commonly applied to data sets with thousands of dimensions such as the term/document frequency matrices that arise in text analysis it can also be applied here. We consider the standard rendering pipeline to be a special case of this operation.

The particular choice of which dimensions to preserve and which to eliminate depends on the particular question being asked by the analyst.

4.2.3 Visualization

Our position on the best overall visualization is that one does not exist. Instead, we advocate the use of multiple linked views (Fig. 1), each of which conveys a different facet of the data. This allows the incorporation of representations already familiar to the analyst (e.g. isobar plots for meteorology) along with standard statistical plots and novel visualization algorithms. We also prefer to display evolution over time by laying out multiple timesteps on screen at once as shown in Figures 1 and 2 instead of by animating a single view. Ensemble data is so complex both visually and conceptually that visual working memory is scarcely ever sufficient to hold the state of the data and its evolution. At the same time we note that there are phenomena where the visual system excels at extracting pattern from chaos via motion. This is especially common in meteorology. Where a still images of Doppler radar may show only rougly congruent blobs of color, an animation immediately reveals the progression of a severe storm as it moves across an area. Similarly, while individual satellite images of a hurricane reveal fascinating structure, a short animation illustrates the entrainment of weather systems for hundreds of miles in all directions.

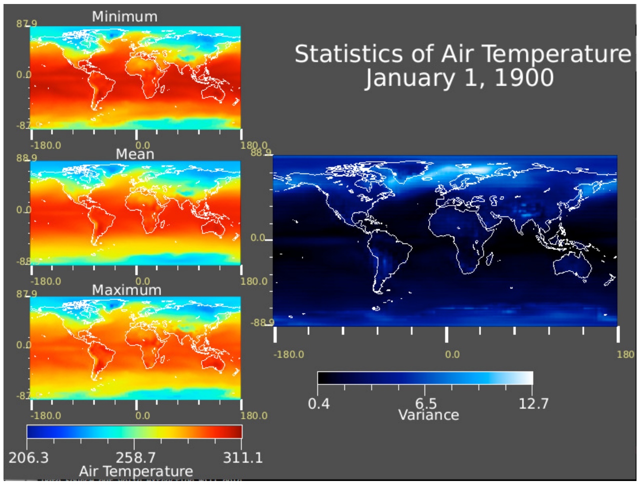


Figure 3: This summary view of global temperature data can be constructed entirely using low-resolution summary statistics. We precompute and cache this summary information in main memory for use as an index and as a way to explore the ensemble's overall behavior.

5. CONCLUSIONS AND OPEN QUESTIONS

Tools for visual analysis and interrogation of large ensemble data sets are still in early development. Many of the efforts to date have been demonstrated on relatively small ensembles that fit comfortably in main memory on a workstation. With the growth of tera- and petascale simulations, however, we may not even be able to load the ensemble on a large parallel machine, let alone a single-user workstation. Rendering all the data is similarly impractical. Instead, we must construct systems and algorithms that combine aspects of databases, simplification, statistical visualization and stream processing to create scalable solutions. In this paper we have highlighted the following pertinent questions:

- How should we store, summarize and access the data?
- What elements or algorithms are common across ensemble analysis in different domains?
- How should we assemble ensemble visualizations?
- How should we reduce a multidimensional ensemble to the two dimensions that can be displayed on a screen or on paper?

We believe that the area of ensemble visualization and analysis is quite young. While some of the major research issues are becoming clear, robust, scalable solutions applicable across a range of application domains are still developing. Although ensemble visualization research is relatively new, ensembles have already established their power as analytical tools in scientific domains of critical importance. For this reason we believe that the need for interactive visual tools for analysis and exploration will remain indefinitely.

6. ACKNOWLEDGMENTS

The authors thank Pat Crossno and Dean Dobranich at Sandia National Labs for valuable discussions and data. We also thank Peer-Timo Bremer and Dean Williams at Lawrence Livermore National Labs for their insight into visualization methods for ensembles of climate data.

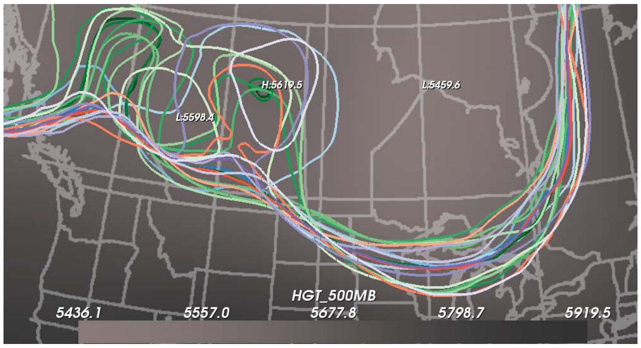


Figure 4: This view requires access to the raw data in order to show the exact position of a particular isocontour in each different ensemble member. Efficient generation of images like this requires index or summary information (to rapidly determine which parts of the raw data must be read) and on-disk organizations that permit efficient random access.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

7. REFERENCES

[1] N. W. S. E. M. Center. Short-range ensemble forecasting. www.emc.ncep.noaa.gov/mmb/SREF/SREF.htm.

[2] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, January 2008.

[3] J. C. Helton. Conceptual and computational basis for the quantification of margins and uncertainty. SAND Report SAND2009-3055, Sandia National Laboratories, June 2009.

[4] A. Love, A. Pang, and D. Kao. Visualizing spatial multivalue data. *IEEE CG & A*, 25(3):69–79, May 2005.

[5] A. Luo, D. T. Kao, J. L. Dungan, and A. Pang. Visualizing spatial distribution data sets. In *Symposium on Visualization (VisSym 2003)*. Eurographics Association, 2003.

[6] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, July 2005.

[7] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. In *Proceedings of ACM SIGMOD (Special Interest Group on Management of Data)*, pages 426–435, 1998.

[8] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, Nov 1997.

[9] V. Pascucci and R. J. Frank. Global static indexing for real-time exploration of very large regular grids. In *Supercomputing '01: Proceedings of the 2001 ACM/IEEE conference on Supercomputing*, pages 2–2, New York, NY, USA, 2001. ACM.

[10] K. C. Potter, A. T. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *Proceedings of the IEEE Workshop on Knowledge*

Discovery from Climate Data: Prediction, Extremes and Impacts, Forthcoming.

[11] T. White. *Hadoop: The Definitive Guide*. O'Reilly Media, 2009.

[12] K. Wu, E. Otoo, and A. Shoshani. Optimizing bitmap indices with efficient compression. *ACM Transactions on Database Systems*, 31:1–38, 2006.

[13] Q. Zhang, J. Liu, and W. Wang. Approximate clustering on distributed data streams. In *IEEE International Conference on Data Engineering*, pages 1131–1139, 2008.