**Annual Report for Period:** 09/2005 - 09/2006
**Submitted on:** 05/29/2006
**Principal Investigator:** Yang, Jun .
**Award ID:** 0238386
**Organization:** Duke University
**Title:**
CAREER: Techniques and Applications of Derived Data Maintenance

## Project Participants

### Senior Personnel

**Name:** Yang, Jun
**Worked for more than 160 Hours:** Yes
**Contribution to Project:**
As the project leader, The PI is responsible for overseeing the direction of the project, supervising a group of gradudate and undergraduate students to do research, and incorporating research into the education curriculum at Duke University. The PI is receiving support for 1.0 summer month in 2006; his travel to the SIGMOD 2006 conference is also partially supported by this CAREER grant.

### Post-doc

### Graduate Student

**Name:** He, Hao
**Worked for more than 160 Hours:** Yes
**Contribution to Project:**
Hao He works with the PI on developing derived data maintenance techniques in the context of XML, graph-structured data, including their application in bioinformatics. In the fall semester of 2005, he is supported by this CAREER grant; his travel to the CIKM 2005 conference was also supported by the CAREER grant.

**Name:** Silberstein, Adam
**Worked for more than 160 Hours:** Yes
**Contribution to Project:**
Adam Silberstein works with the PI on developing query processing techniques for wireless sensor networks. In 2005-2006, he is primarily supported by PI's other funding sources; his travel to the ICDE 2005 conference was supported by this CAREER grant.

**Name:** Xie, Junyi
**Worked for more than 160 Hours:** Yes
**Contribution to Project:**
Junyi Xie works with the PI on developing derived data maintenance techniques in the context of view maintenance, stream and continuous query processing. In the spring semester of 2006, he is supported by this CAREER grant.

**Name:** Chandramouli, Badrish
**Worked for more than 160 Hours:** Yes
**Contribution to Project:**
Badrish Chandramouli works with the PI on developing derived data maintenance techniques in the context of network data querying. During the 2005-2006 academic year, his research assistantship is supported

by this CAREER grant.

**Undergraduate Student**

**Technician, Programmer**

**Other Participant**

**Research Experience for Undergraduates**

**Name:** Wu, Congyi

**Worked for more than 160 Hours:** No

**Contribution to Project:**

Congyi Wu is an undergraduate student working with the PI and Hao He on applying derived data maintenance techniques to bioinformatics. In summer 2006, his summer research internship is support by an REU supplement to this CAREER grant. His internship started on May 10, 2006, and will have worked for 400 hours by the end of summer 2006.

**Years of schooling completed:** Freshman

**Home Institution:** Same as Research Site

**Home Institution if Other:**

**Home Institution Highest Degree Granted(in fields supported by NSF):** Doctoral Degree

**Fiscal year(s) REU Participant supported:** 2005

**REU Funding:** No Info

**Name:** Filpus, Gregory

**Worked for more than 160 Hours:** No

**Contribution to Project:**

Greg Filpus is an undergraduate student working with the PI and Adam Silberstein on developing query processing techniques for wireless sensor networks. In summer 2006, his summer research internship is support by an REU supplement to this CAREER grant. His internship started on May 15, 2006, and will have worked for 400 hours by the end of summer 2006.

**Years of schooling completed:** Freshman

**Home Institution:** Same as Research Site

**Home Institution if Other:**

**Home Institution Highest Degree Granted(in fields supported by NSF):** Doctoral Degree

**Fiscal year(s) REU Participant supported:** 2005

**REU Funding:** No Info

## Organizational Partners

## Other Collaborators or Contacts

Pankaj K. Agarawal, Carla S. Ellis, and Kamesh Munagala. Faculty members in the Department of Computer Science, Duke University.

Rebecca L. Braynard, Ke Yi, and Hai Yu. Students in the Department of Computer Science, Duke University.

Yuguo Chen. Faculty member in the Department of Statistics, University

of Illinois at Urbana-Champaign.

Amin Vahdat. Faculty member in the Computer Science Department, University of California at San Diego.

James S. Clark. Faculty member in the Duke University School of Environment.

Cliburn Chan, Lindsay G. Cowell, and Thomas B. Kelper. Faculty members at the Center for Bioinformatics and Computational Biology, Duke University.

David F. Kong. Duke University Medical Center.

Ioana Stanoi, Haixun Wang, and Philip S. Yu. IBM T. J. Watson Research Center.

Jeffrey Yu Xu. Faculty member at the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong.

## Activities and Findings

**Research and Education Activities:**

The focus of this CAREER project is on techniques and applications of derived data maintenance. Derived data is the result of applying some transformation, structural or computational, to base data. The use of derived data to facilitate access to base data is a recurring technique in many areas of computer science. Used in hardware and software caches, derived data speeds up access to base data. Used in replicated systems, it improves reliability and performance of applications in a wide-area network. Used as index structures, it provides fast alternative access paths to base data. Used as materialized views in databases or data warehouses, it improves the performance of complex queries over base data. Used as synopses, it provides fast, approximate answers to queries or statistics needed for cost-based optimization. Derived data may vary in complexity: it can be a simple copy of base data, in the cases of caching and replication, or it can be the result of complex transformations, in the cases of indexes and materialized views. Derived data may also vary in accuracy: caches and materialized views are usually exact, while synopses are approximate. Regardless of the varying forms, purposes, complexity, and accuracy of derived data, it must be maintained when base data is updated. Thus, derived data maintenance is a fundamental problem in computer science. It is also an evolving problem: existing techniques are constantly challenged by the explosive growth in data volume and number of data producers and consumers, and by increasing diversity in data formats and storage and communication media. Traditionally, derived data maintenance has been tackled separately in different contexts, e.g., index updates and materialized view maintenance in databases, cache coherence and replication protocols in distributed systems. Although they share the same underlying theme, these techniques have been developed and applied largely disjointly. Newer and more complex data management tasks, however, call for creative combinations of the traditionally

separate ideas. Semantic caching, which has received tremendous interests recently for its applications in caching dynamic Web contents, is a good example of incorporating the idea of materialized views into a cache. With ``outside-the-box'' thinking such as semantic caching, we seek to discover more techniques that combine multiple flavors of derived data to provide better solutions to problems.

In Year 3 of this project, we have investigated the following specific research problems:

1. Derived data in scalable continuous query processing. Continuous query processing has attracted much interest from the database community recently because of its wide range of traditional and emerging applications, e.g., trigger and production rule processing, data monitoring, stream processing, and publish/subscribe systems. In contrast to traditional query systems, where each query is run once against a snapshot of the database, continuous query systems support standing queries that continuously generate new results (or changes to results) as data updates continue to arrive in a stream. In this sense, continuous query processing has much in common with incremental view maintenance, and can be regarded also as a problem of derived data maintenance. One of the main challenges in continuous query processing is how to handle a large number of continuous queries in a scalable way. For each incoming data update, the system needs to identify the subset of continuous queries whose results are affected by the data update, and compute changes to these results. If there are many continuous queries, a brute-force approach that processes each of them in turn will be inefficient and unable to meet the response-time requirement of most target applications. One important insight gained by research on scalable continuous query processing is the interchangeable roles played by queries and data. In continuous query systems, continuous queries can be treated as data, while each data update can be treated as a query requesting the subset of continuous queries affected by the update. Thus, it is natural to apply indexing and query processing techniques traditionally intended for data to continuous queries. Most existing work on indexing continuous relational queries has focused on selections. As far as we know, there has been little work on how to process more complex continuous queries (e.g., joins) scalably. We have been developing efficient processing techniques in collaboration with Pankaj K.  Agarwal, a colleague in Duke Computer Science. In particular, we have developed novel, 'input-sensitive' schemes for indexing continuous joins with range conditions. These schemes exploits the clusteredness of the range conditions being indexed: More clustered queries lead to more efficient processing. We have also obtained other results including lower bounds on the inherent complexity of the problem, and data structures with space-time tradeoffs. Some of our more theoretical results have been published in ISAAC 2005, while more practical extensions and implementation and experimentation results are currently under submission.

2. Derived data in wide-area network applications. We have explored the use of derived data in two applications. (1) Wide-area network querying. As networks continue to grow in size and complexity, distributed network monitoring and resource querying are becoming increasingly difficult. Our aim is to design, build, and evaluate a scalable infrastructure for answering queries over distributed

measurements, using reduced costs (in terms of both network traffic and query latency) while maintaining required precision. To this end, we use bounded approximate caches, a form of derived data. Each network node owns a set of numerical measurements and actively maintains bounds on these values cached at other nodes. We can answer queries approximately, using bounds from nearby caches to avoid contacting the owners directly. We focus on developing efficient and scalable techniques to place, locate, and manage caches across a large network. We have developed two approaches: One uses a recursive partitioning of the network space to place caches in a static, controlled manner, while the other uses a locality-aware distributed hash table to place caches in a dynamic and decentralized manner. Experiments using large-scale network emulation show that our techniques are very effective in reducing query costs while generating an acceptable amount of background traffic; they are also able to exploit various forms of locality that are naturally present in queries, and adapt to volatility of measurements. The work has been published in DASFAA 2006. (2) Wide-area publish/subscribe. Subscriptions are essentially derived data that must be maintained as new publish messages arrive. The work performed by a publish/subscribe system can conceptually be divided into subscription processing and notification dissemination. Traditionally, research in the database and networking communities has focused on these aspects in isolation. The interface between the database server and the network is often overlooked by previous research. At one extreme, database servers are directly responsible for notifying individual subscribers; at the other extreme, updates are injected directly into the network, and the network is solely responsible for processing subscriptions and forwarding notifications. These extremes are unsuitable for complex and stateful subscription queries. We explored the design space between the two extremes, and devised solutions that incorporate both database-side and network-side considerations in order to reduce the communication and server load and maintain system scalability. Our techniques apply to a broad range of stateful query types. Experiments with link-level network simulation show that by exploiting the query semantics and building an appropriate interface between the database and the network, it is possible to achieve orders-of-magnitude savings in network traffic at low server-side processing cost. This work has been published in SIGMOD 2006.

3. Derived data for graph reachability. Graph-structured data has found a growing number of important applications recently. In bioinformatics, protein interactions, metabolic pathways, and gene regulatory networks are modeled as directed graphs. In Semantic Web, two key technologies, RDF and OWL, are designed to capture graph data. In addition, although XML is generally modeled as a tree, many XML applications treat cross-reference edges (IDREF/ID) as first-class citizens, making the data graph-structured. Testing reachability between two nodes in a graph is a fundamental operation crucial to many graph data processing tasks. General graph reachability has been a well-known difficult problem, and existing solutions often fail to meet the requirement of new applications. A promising approach is to derived data precomputed from the graph as an index. In our work published in CIKM 2005, we observed that many graphs in our target applications exhibit structures (or substructures) that make the problem amenable to more efficient indexing. We developed HLSS

(Hierarchical Labeling of Sub-Structures), which identifies different types of substructures within a graph and encodes them using techniques suitable to the characteristics of each of them. Experiments show that HLSS handles different types of graphs well, while existing approaches, such as interval-based labeling and 2-hop labeling, fall prey to graphs with substructures they are not designed to handle. In our work published in ICDE 2006, we further observed that in many application scenarios the data graphs are extremely sparse---the edge-to-node ratio is rarely higher than 2, and usually very close to 1; examples include XMark and a number of bioinformatics datasets. We developed a labeling scheme that supports reachability queries in constant time, while consuming space linear in the size of the graph and quadratic in the number of non-tree edges, which is small for sparse graphs. Experiments show that our approach is much more efficient than state-of-the-art approaches such as 2-hop.

4. Derived data in wireless sensor networks. In collaboration with Duke University School of Environment, we are building a wireless sensor network in Duke Forest to study how various environmental variables influence forest growth. Wireless sensor networks are capable of generating a vast amount of data; this data, however, must be sparingly extracted to conserve energy, usually the most precious resource in battery-powered nodes. Our collaborative team is developing novel OS, networking, and data service layers that implement a dynamic, data-driven approach to energy-efficient sensing and communication. While the majority of this effort has been recently funded under NSF's DDDAS program in January 2006, we have been actively studying the use of derived data in sensor data processing through the support of this CAREER grant, and have obtained a series of results. (1) In our work published in ICDE 2006, we considered snapshot top-k queries in sensor When approximation is acceptable, a model-driven approach to query processing is effective in saving energy by avoiding contacting nodes whose values can be predicted or are unlikely to be in the result set. Models are typically derived from past sensor readings. To optimize queries such as top-k, however, reasoning directly with models of joint probability distributions can be prohibitively expensive. Instead of using models explicitly, we propose to use samples of past sensor readings, which are much a simpler form of derived data. Not only are such samples easy to maintain, but they are also computationally efficient in query optimization. With these samples, we can formulate the problem of optimizing approximate top-k queries under an energy constraint as a linear program, allowing incorporation of advanced features such as topology-aware planning and in-network filtering. (2) In one of our SIGMOD 2006 papers, we developed energy-efficient algorithms for continuously monitoring extreme values in a sensor network. An extreme-value query is essentially a top-1 query, and our result can be extended to top-k queries. Our algorithms employ a hierarchy of local constraints, or thresholds, to leverage network topology such that message-passing is localized. These local constraints can be seen as data (more precisely, approximate bounds) derived from the current state of the network, and need to be maintained whenever they are violated. Our experiments show that such derived data, appropriately designed and placed, can lead to dramatic energy savings. (3) In another one of our SIGMOD 2006 papers (an earlier version of which also appeared as an ICDE 2006 poster), we tackled the problem of simply monitoring all sensor readings at all times. It turns out that

derived data, in the form of spatio-temporal models of sensor readings, can be used to reduce the monitoring cost. Inside the network, nodes employs a collection of simple models to decide whether to report new readings---reporting is suppressed if a new value does not differ from one predicted by the model. Outside the network, the base station maintains replicas of these models and use them to infer current values of those readings that are not reported. While previous work has explored suppression based on temporal and spatial models separately, we demonstrate how to combine both temporal and spatial suppression effectively using a technique called constraint chaining. Constraint chaining builds a network of constraints that are maintained locally, but allow a global view of values to be maintained with minimal communication cost.

5. Asymmetric batch incremental view maintenance. We continued our investigation of asymmetric batch incremental view maintenance from Year 2. Incremental view maintenance is probably the most well-studied problem about derived data in the database community. It has a growing number of applications recently, including for example data warehousing and publish/subscribe systems. Batch processing of base table modifications, when applicable, can be much more efficient than processing individual modifications one at a time. We tackle the problem of finding the most efficient batch incremental maintenance strategy under a refresh response time constraint; that is, at any point in time, the system, upon request, must be able to bring the view up to date within a specified amount of time. The traditional approach is to process all batched modifications relevant to the view whenever the constraint is violated. However, we observe that there often exists natural asymmetry among different components of the maintenance cost; for example, modifications on one base table might be cheaper to process than those on another base table because of some index. We exploit such asymmetries using an unconventional strategy that selectively processes modifications on some base tables while keeping batching others. We demonstrated in an ICDE 2005 paper that this strategy offers substantial performance gains over traditional deferred view maintenance techniques. However, we assumed some knowledge of the arrival pattern of base table updates in order to obtain a provably good maintenance plan. In Year 3, we have developed more robust online algorithms that are competitive against any oblivious adversary, thereby lifting our previous assumption. This work was published in ESA 2005.

The progress we have made in Year 3, as summarized above, are in line with the modified research plan outlined in the project report from Years 1 and 2. We have broaden our study of derived data maintenance to continuous queries in a networked setting, including wide-area publish/subscribe and wireless sensor networks. In parallel, we continue to work on derived data in the forms of views and indexes for traditional databases and graph-structured data. In Year 4, we plan to (1) continue our work on scalable subscription processing and notification in the context of a wide-area publish/subscribe system with a rich subscription language; (2) continue our work on indexing graph-structured data; (3) work with our bioinformatics collaborators to apply our techniques (for both subscription processing and graph indexing) to manage lineage, dependency, and versioning of derived datasets that arise in computational biology workflows; (4) work with our collaborators in Duke University School of Environment to apply

derived data techniques to wireless sensor networks. We are currently seeking additional funding to complement our work in the area of wide-area publish/subscribe systems.

In terms of educational activities, I have continued to incorporate current research topics into both undergraduate and graduate database course at Duke University. The undergraduate database course I offered in Fall 2005 covered a substantial amount of material drawn from the latest research. Currently, I am working with undergraduate summer interns (funded through an REU supplement to this CAREER grant), both of whom are working on application areas of derived data: One is working on query processing in wireless sensor networks, while the other is working on a system for managing lineage, dependency, and versioning of derived datasets in computational biology workflows.

**Findings:**

We have made significant progress in studying the derived data maintenance problem in multiple application domains, including view maintenance, data warehousing, stream data processing, XML and graph indexing, continuous query processing in wide-area networks and sensor networks. Published results from this grant so far include:

Traditional settings:
* An efficient method for top-k view maintenance that incorporates the idea of caching (ICDE 2003).
* A new approach to batch incremental view maintenance that exploits asymmetry in maintenance cost components (ICDE 2005 and ESA 2005).

XML and graph-structured data:
* A novel XML structural index (ICDE 2004) utilizing derived data at multiple resolutions.
* Efficient incremental maintenance algorithms for XML structural indexes (SIGMOD 2004), which incorporate the use of auxiliary data.
* Efficient maintenance of order-based labeling for dynamic XML documents, with different degrees of materialization to provide a tradeoff between query and update performance (ICDE 2005).
* A hybrid labeling scheme for graph reachability that identifies different types of substructures within a graph and encodes them using techniques suitable to the characteristics of each (CIKM 2005).
* A labeling scheme supporting constant-time graph reachability queries while remaining space-efficient for sparse graphs (ICDE 2006).

Wide-area network querying:
* A system for distributed network monitoring and resource querying by intelligently placing, locating, and managing bounded approximated caches across the network (DASFAA 2006).

Continuous query processing in stream and publish/subscribe systems:
* Framework and techniques for managing the state of a stream join to maximize result completeness, which is related to the classic caching problem (SIGMOD 2005).
* New, input-sensitive approaches to scalable processing of continuous join queries (ISAAC 2005, and technical report submitted for publication).
* A new approach towards wide-area publish/subscribe that examines the spectrum of possibilities of interfacing subscription processing and

notification delivery for more efficient support of stateful subscriptions (SIGMOD 2006).

Applications of derived data in sensor networks:
* A model-driven approach to snapshot top-k queries that uses samples of past sensor readings and linear programming for optimization (ICDE 2006).
* Energy-efficient algorithms for continuously monitoring extreme values using a hierarchy of local constraints, or thresholds (SIGMOD 2006).
* Energy-efficient monitoring using spatio-temporal suppression and a chain of locally monitored constraints for reconstructing the global view (poster paper in ICDE 2006; full paper in SIGMOD 2006).

We are actively working with our collaborators in bioinformatics and ecology and applying the above results to real-world problems. Based on these findings we believe that the direction we are currently pursuing is a promising one. For detailed descriptions of these findings please refer to the section of this report on research and education activities.

**Training and Development:**

The PI has advised the following students in the context of this project:

Ph.D. students: Hao He, Adam Silberstein, Junyi Xie, Badrish Chandramouli.

M.S. students: Zhihui Wang (thesis completed in 2003), Wenbin Pan (thesis completed in 2004).

Undergraduate students: Christopher N. Bond (BS with High Distinction, 2005), Gregory Filpus, Congyi Wu.

**Outreach Activities:**

The PI has been active in running the Carolina Database Research Group (http://www.cs.duke.edu/cdb/) with a group of database researchers in North Carolina, including members from Duke, North Carolina State University, University of North Carolina at Chapel Hill, Charlotte, and Greensboro. We hold monthly meetings and are currently running a seminar series, which have been a great resource for facilitating student and faculty interaction across institutions and attracting student interests in database research. The PI was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

## Journal Publications

## Books or Other One-time Publications

Ke Yi, Hai Yu, Jun Yang, Gangqiang Xia, and Yuguo Chen, "Efficient Maintenance of Materialized Top-k Views", (2003). Proceedings, Published
Collection: Proceedings of the 19th International Conference on Data Engineering (ICDE '03)
Bibliography: Bangalore, India

Junyi Xie, Jun Yang, and Yuguo Chen, "On Joining and Caching Stochastic Streams", (2005). Proceedings, Published
Collection: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data
Bibliography: Baltimore, Maryland, June 2005

Adam Silberstein and Jun Yang, "NeXSort: Sorting XML in External Memory", (2004). Proceedings, Published
Collection: Proceedings of the 20th International Conference on Data Engineering (ICDE '04)
Bibliography: Boston, Massachusetts, March 2004

Hao He and Jun Yang, "Multiresolution Indexing of XML for Frequent Queries", (2004). Proceedings, Published
Collection: Proceedings of the 20th International Conference on Data Engineering (ICDE '04)
Bibliography: Boston, Massachusetts, March 2004

Ke Yi, Hao He, Ioana Stanoi, and Jun Yang, "Incremental Maintenance of XML Structural Indexes", (2004). Proceedings, Published
Collection: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD '04)
Bibliography: Paris, France, June 2004

Zhihui Wang, "Multiple-View Maintenance with Semantic Caching", (2003). Thesis, Published
Collection: M.S. Thesis, Duke University
Bibliography: Durham, North Carolina, August 2003

Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu, "On Scalable Processing of Continuous Joins", (2004). Technical Report, Submitted
Collection: Technical Report, Department of Computer Science, Duke University
Bibliography: Durham, North Carolina, December 2004

Adam Silberstein, Hao He, Ke Yi, and Jun Yang, "BOXes: Efficient Maintenance of Order-Based Labeling for Dynamic XML Data", (2005). Proceedings, Published
Collection: Proceedings of the 21st International Conference on Data Engineering
Bibliography: Tokyo, Japan, April 2005

Hao He, Junyi Xie, Jun Yang, and Hai Yu, "Asymmetric Batch Incremental View Maintenance", (2005). Proceedings, Published
Collection: Proceedings of the 21st International Conference on Data Engineering
Bibliography: Tokyo, Japan, April 2005

Kamesh Munagala, Jun Yang, and Hai Yu, "Online View Maintenance Under a Response-Time Constraint", (2005). Proceedings, Published
Collection: Proceedings of the 13th Annual European Symposium on Algorithms (ESA '05)
Bibliography: Mallorca, Spain, October 2005

Hao He, Haixun Wang, Jun Yang, and Philip S. Yu, "Compact Reachability Labeling for Graph-Structured Data", (2005). Proceedings, Published
Collection: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)
Bibliography: Bremen, Germany, November 2005

Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu, "Monitoring Continuous Band-Join Queries over Dynamic Data", (2005). Proceedings, Published
Collection: Proceedings of the 16th Annual International Symposium on Algorithms and Computation (ISAAC '05)
Bibliography: Sanya, Hainan, China, December 2005

Adam Silberstein, Rebecca Braynard, and Jun Yang, "Energy-Efficient Continuous Isoline Queries in Sensor Networks (Poster Paper)", (2006). Proceedings, Published
Collection: Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)
Bibliography: Atlanta, Georgia, USA, April 2006

Haixun Wang, Hao He, Jun Yang, Philip S. Yu, and Jeffrey Xu Yu, "Dual Labeling: Answering Graph Reachability Queries in Constant Time", (2006). Proceedings, Published

Collection: Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)
Bibliography: Atlanta, Georgia, USA, April 2006

Adam Silberstein, Rebecca Braynard, Carla Ellis, Kamesh Munagala, and Jun Yang, "A Sampling-Based Approach to Optimizing Top-k Queries in Sensor Networks", (2006). Proceedings, Published
Collection: Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)
Bibliography: Atlanta, Georgia, USA, April 2006

Badrish Chandramouli, Jun Yang, and Amin Vahdat, "Distributed Network Querying with Bounded Approximate Caching", (2006). Proceedings, Published
Collection: Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA '06)
Bibliography: Singapore, April 2006

Adam Silberstein, Kamesh Munagala, and Jun Yang, "Energy-Efficient Monitoring of Extreme Values in Sensor Networks", (2006). Proceedings, Published
Collection: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)
Bibliography: Chicago, Illinois, USA, June 2006

Adam Silberstein, Rebecca Braynard, and Jun Yang, "Constraint-Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks", (2006). Proceedings, Published
Collection: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)
Bibliography: Chicago, Illinois, USA, June 2006

Badrish Chandramouli, Junyi Xie, and Jun Yang, "On the Database/Network Interface in Large-Scale Publish/Subscribe Systems", (2006). Proceedings, Published
Collection: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)
Bibliography: Chicago, Illinois, USA, June 2006

## Web/Internet Site

**URL(s):**
http://www.cs.duke.edu/dbgroup/ddm/
**Description:**

## Other Specific Products

## Contributions

**Contributions within Discipline:**
We have made contributions to multiple application domains of derived data maintenance, including view maintenance, data warehousing, stream data processing, XML and graph indexing, continuous query processing in wide-area networks and sensor networks. A number of the contributions have been published in premier conferences (five full papers in SIGMOD 2004-2006, six full papers in ICDE 2003-2006, one paper each in CIKM 2005, ESA 2005, ISAAC 2005, and DASFAA 2006). For detailed descriptions of these contributions please refer to the section of this report on research and education activities.

In addition to serving on numerous program committees, the PI has been active in running the Carolina Database Research Group, and was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

**Contributions to Other Disciplines:**

The PI has been actively applying derived data techniques to areas beyond computer science. Specifically, the PI has been working with a group of computational immunologists led by Dr. Thomas B. Kepler at Duke University on developing a system for tracking lineage, dependency, and versioning of derived datasets in computational biology workflows. Also, the PI has been collaborating with a group of ecologists led by Dr. James S. Clark at the Duke University School of Environment on developing a wireless sensor network in Duke Forest to study how various environmental variables influence forest growth.

**Contributions to Human Resource Development:**

**Contributions to Resources for Research and Education:**

The PI has been active in running the Carolina Database Research Group (http://www.cs.duke.edu/cdb/) with a group of database researchers in North Carolina, including members from Duke, North Carolina State University, University of North Carolina at Chapel Hill, Charlotte, and Greensboro. We hold monthly meetings and are currently running a seminar series, which have been a great resource for facilitating student and faculty interaction across institutions and attracting student interests in database research. The PI was one of organizers of the First Southeast Workshop on Data and Information Management in March 2006.

**Contributions Beyond Science and Engineering:**

## Special Requirements

**Special reporting requirements:** None

**Change in Objectives or Scope:** None

**Unobligated funds:** less than 20 percent of current funds

**Animal, Human Subjects, Biohazards:** None

## Categories for which nothing is reported:

Organizational Partners

Any Journal

Any Product

Contributions: To Any Human Resource Development

Contributions: To Any Beyond Science and Engineering