# A PROTOTYPE INTEGRATED TRANSACTION DATA ANALYSIS AND VISUALIZATION ENVIRONMENT FOR THE TRANSPORTATION, DISTRIBUTION AND LOGISTICS SECTOR

A Proposal for Exploratory Research Submitted to the
E-Enterprise Center at Discovery Park by

**Chris Clifton**
Associate Professor
Department of Computer Sciences
Indiana Center for Database Systems

**Ananth Iyer**
Professor
Krannert Graduate School of Management
Laboratory for Extended Enterprises at Purdue (LEEAP)

**Reha Uzsoy**
Professor
School of Industrial Engineering
Laboratory for Extended Enterprises at Purdue (LEEAP)

August 2, 2002

**Introduction**

Today's leading companies are well aware of the importance of managing and designing the entire supply chain effectively. The increasingly rapid pace of technological change, particularly in information technology, has led to unprecedented new opportunities such as electronic commerce whose implications are not yet fully understood. Management is more and more frequently forced to make decisions for which little or no precedent exists and whose associated risks and gains are difficult to assess.

While these developments in information technology offer tremendous opportunities for companies to improve their operations, we believe that the decision technologies that take this transactional data as input and use it to develop better decisions that improve company performance have lagged behind. Today decision makers are faced with massive streams of data of differing nature, reliability and level of detail. For example, the ERP systems sold in the market today reflect the latest information technology in distributed databases, Internet connectivity and user interfaces. However, the supply chain management engines in many of them are direct outgrowths of the Material Requirements Planning paradigm developed in the late 1960s. Much of the advanced planning and scheduling software sold in the market is based on research in artificial intelligence-based scheduling from the mid 1980s. While research in supply chain management has attempted to include game theoretic issues and contracts as part of decision making, most models have focused on single period environments or at best a finite number of periods to permit model tractability. The growth of data mining technology provides tools to analyze these massive streams of data, but application of these tools to supply chain problems has been limited. In our opinion, there is a significant gap between the richness of the data available to today's business enterprises and their ability to leverage this data to make better decisions.

As an example of the type of problem we hope to address, imagine several small trucking companies that wish to collaborate to determine if they can share capacity. One obvious solution is back-hauling: If Great Eastern is swamped with deliveries from Ft. Wayne to Gary, and Western Trucking is overloaded with business from Chicago to South Bend, they can collaborate to fill their otherwise empty trucks on the return trip. At first glance, this problem seems easily modeled – represent routes as a graph, and search for overlapping edges. However, this only captures one type of collaboration – what if the routes are instead Ft. Wayne to Evansville, and

Gary to Cincinnati? There are no back-hauling opportunities, but the junction of routes in Indianapolis opens possibilities shared depot and warehousing facilities. What we need is data mining techniques that highlight the interesting points in the graph (cycles, junctions, ?) so domain experts can explore the potential for collaboration.

Data mining techniques for such graph analysis exist (at least in the research lab). However, lack of interaction with domain experts has left a gap between the problem and the technology: Time. The temporal nature of this problem is crucial – if both trucking companies converge on Indianapolis at the same time, a shared warehouse would be inappropriate.

The problem described above is one instance of the general problem which is essentially "how should fragmented *temporal* pockets of capacity across independent entities be aggregated into a coherent economically viable supply system that can satisfy demand." As such, a collaboration tool that permits the aggregation of such capacity across independent entities would provide a crucial link to create such a viable system. In addition, the ability to visualize such capacity streams to develop a model of their structure would also be imperative. Solving this problem will impact other industries as well e.g., custom tool and die shops, foundry wafer fabricators, etc. also have potential for this type of collaboration.

We propose to develop a facility with transactional data challenge problems and data sets to drive research in this area. It will provide an environment where domain knowledge, modeling capability, and computational expertise come together to identify and solve new problems. Examples such as the University of California at Irvine Repository of Machine Learning Databases and the Linguistic Data Consortium's collection of language-related corpuses demonstrate how advances in information research can be driven by the availability of focused challenge problems and data sets.

Based on our observations of a variety of industrial sectors, we believe that the formulation of meaningful problems in the domain of the E-Enterprise to the point of building a generalizable, transferable base of knowledge for their solution requires a novel approach combining the mathematical and computer modeling tools from disciplines that have had only limited interaction in the past, such as industrial engineering, operations research, computer science, along with the an understanding of the business context for decision making. Thus there is a need to develop tools, facilities, and methodologies that facilitate the interaction of academic specialists in the related disciplines with industrial researchers and practitioners familiar with the

issues in the field. The development of such an environment would serve as a tremendous catalyst to both applied and scholarly research, as well as provide a leading-edge educational facility for the new generation of business and technical professionals that must operate in the E-Enterprise of today.

**Methodology**

Under this proposal we will develop the first stage of a Transactional Data Analysis and Visualization Laboratory (TDAVL). The Laboratory will consist of a hardware and software platform in which we can analyze transactional data sets from industrial partners and use them as the basis for research in how to manage the date and provide decision support in today's data rich environment, along with a methodology for interaction to address these problems. In order to focus the project in the short term, we propose to concentrate on a particular industrial sector, that of Transportation, Distribution and Logistics (TDL), and a specific set of management issues, that of constructive collaboration between firms.

Based on research done in 2001-2 by the Battelle Group for the Central Indiana Corporate Partnership, this sector constitutes a significant fraction of economic activity in the state. Many companies in this sector with operations in Indiana, such as Fedex, United Parcel Service and North American Logistics, have significant data collection and tracking capabilities that would provide meaningful data sets to begin studying. It is likely that even using data sets that are several years old would be enough to provide a useful basis from which to begin our studies.

One problem with obtaining sample transactional datasets is privacy and corporate secrecy. A possible solution is to randomize the data, altering it so that actual transactions are no longer revealed. Emerging research in privacy preserving data mining provides techniques for learning valid models from such data. This option will ease the task of obtaining data sets that would otherwise be unavailable for research use.

The issue of collaboration is also especially meaningful in this domain. While some of the larger companies such as those mentioned above have developed significant information technology skills that provide a major competitive advantage, there are a large number of small to medium sized firms involved in this sector that do not have the resources to develop their own technology solutions, and engage in cutthroat competition for commoditized services that

prevent them from developing further. Given the widely documented trend towards outsourcing of TDL processes by both large and small companies in a wide range of industries, it is easy to envision situations in which a number of these companies could collaborate by sharing resources and information to provide better customer service at lower cost, opening the possibility of new business models and a much more vibrant, economically viable sector. This offers the potential role for the mediated use of the TDAVL environment to explore collaboration across such businesses.

Collaboration in a free market is problematic: collaborators can also be competitors, even in the scope of the same project. We have developed techniques for data mining in such an environment: learning high-level relationships and patterns from a combined dataset, without revealing the actual data to any but the "owning" party. This project will enable application of this work to the transportation, distribution, and logistics domain. Interaction with companies and domain experts will also lead to new challenge problems, driving basic research in this area.

Our focus on transactional data has several motivations. First, transactional data are collected in some form and level of detail by all but the smallest firms, and the information technology infrastructure for managing this information at its basic level is well understood. Second, such data allows us to examine how different threads of events, such as shipments to a particular geographic location or a particular customer, evolved over time, offering the possibility of reasoning about the implicit and explicit decision processes being used. Finally, although transaction data represent events that have actually happened in the past, most planning systems oriented towards future activities draw on this data to make inferences about future conditions.

The focus on collaboration is motivated by the widely recognized potential of positive collaboration among firms to yield significant improvements in cost and service, and the equally well recognized lack of a common view of just what collaboration should be. Based on our discussions with industrial partners in the Laboratory for Extended Enterprises at Purdue (LEEAP) and with software companies, most of the software tools being developed in the market today are for sharing specific data items, not shared decision making. While electronic marketplaces and auction schemes have been proposed and studied as one possible mode of collaboration, we feel that this is an unnecessarily limiting view of collaboration. A far more interesting question, as evidenced by the increasing trend in industry towards long-term strategic

partnerships, is how firms can structure a long-term collaborative relationship with a specific objective in a mutually beneficial manner.

Our basic methodology within the scope of this exploratory proposal will be to work closely with existing industrial partners to develop an architecture for the proposed environment. In order to provide a concrete basis for these discussions, we propose to build a limited prototype of this environment using one or two data sets and illustrate how these data sets can be used to study one, or maybe two, very specific issues in collaboration in the TDL sector. Our team has already been involved in the development of several such prototypes, such as the Supply Chain Optimization and Protocol Environment (SCOPE) project, funded by the National Science Foundation, and LEEAP, where Professors Uzsoy and Iyer are active participants. Another such prototype, in this case of an electronic marketplace for small companies such as custom tool and die makers to share production capacity, has been developed by Professor Iyer and his students. Professor Iyer is also working on a project (joint with Professor Deshpande) that will use detailed transactional data from a spare part management facility. Such a project could make substantial use of the TDVAL environment. Professor Clifton's group has developed methods for data mining in distributed environments, specifically where security and privacy concerns limit the sharing of data.

**Nature of Interdisciplinary Interaction**

The development of the proposed environment for using transactional data sets to study collaboration lies at the intersection of the areas of decision support and modeling in supply chains, and the management and mining of large data sets that are rich in contextual information. A key issue is in determining the right level of aggregation at which transactional data can be viewed by internal decision support systems and shared with collaborating firms. Another central issue here is that of allowing data to be shared in a manner that protects the parties involved from possible hostile acts and gaming by malevolent partners. This may well require the development of data mining and decision making procedures in which none of the parties is privy to detailed information, and can only have access to information that does not allow them to infer potentially damaging information about other partners that they can use to their own advantage. Hence the interdisciplinary element in this project lies in marrying the data management and data security skills of Professor Clifton's group with the quantitative decision modeling skills, supply

chain domain expertise and industrial partners that Professors Uzsoy and Iyer have developed through their work over the last three years in the Laboratory for Extended Enterprises at Purdue (LEEAP). This project also has the potential to build closer linkages between LEEAP and a number of the active centers based in the School of Science, such as the Indiana Center for Database Systems (ICDS) and the Center for Education and Research in Information Assurance and Security (CERIAS).

We expect the problem solving approach and interdisciplinary interaction to follow the following model:

1. Domain experts and scientists will work together to identify a challenge problem. Domain experts may include representatives from industry, although initially we expect faculty from the schools of Management and Industrial Engineering to use their domain knowledge to fulfill that role. An example challenge problem would be "identify potential for collaboration among trucking companies based on pockets of available capacity".
2. Obtain a data set to serve as a base for study. Where possible, industrial contacts will be used to obtain real data. We will also develop facilities to construct simulated data sets – we anticipate that the initial demonstration project will use simulated data. Again, this will demand close cooperation between LEEAP (knowledge of the type of transactional data that could potentially be available) and the School of Science (constructing and storing the data sets).
3. Develop approaches to characterize objects in such a data domain and develop tools to visualize the data and thus generate insight into the data characteristics. The visual description of the spatial data will permit management insights to alternate approaches to bundle the temporal capacity to develop economic viable supply. In addition, the feasible use of the capacity may well require agreements across independent firms which might require creative inputs from managers. We believe that a graphical interface can enable us to develop a mathematical characterization of the problem environment. This will be useful as we do the next step. In addition, the role of faculty and graduate students as mediators in such an environment justifies its potential location at Purdue University.

4. Interact to mathematically characterize challenge problem in terms of data set. The goal of this step is to identify a mathematical model whose solution would not only address the challenge problem in terms of the test data set (trucking collaboration), but would extend to a variety of related problems (e.g., custom tool and die production). This may require modifying/extending the data set and/or challenge problem. One approach will be to imagine possible results such as the back-hauling example described previously, modify the data to enable discovery of the result, and then develop a model that could expose that result.
5. The participants will apply their expertise to:
    a. Develop data mining techniques that construct the model from the data, and
    b. Identify potential results beyond those from step 4, and create data set extensions to test the data mining algorithms.

    These two tasks demand different expertise, and separating them will help ensure that the results generalize beyond the specific possible results identified. However, it is also likely that both tasks will identify limitations in the model, requiring a return to step 4. These steps will be iterated – increasing the understanding of data mining techniques among domain experts and vice-versa – on a continuous basis.
6. Test the developed techniques, and disseminate the results to the relevant research communities.

**Potential for Follow-on Funding**

We would like to explore the possibility of a proposal to the Central Indiana Corporate Partnership to carry out initial study in the TDAV Lab at Purdue. There are a wide variety of agencies, such as DARPA, NSF and the Federal and State Departments of Transportation, where specific research proposals that emerge could be submitted. This type of activity would also support a number of ongoing externally funded projects such as the SCOPE project by LEEAP, the current COAST project by Professors Iyer and Deshpande, pedagogical tools to explore the impact of sharing capacity across competitors (developed by Professors Iyer and Ward) and basic research in privacy preserving data mining funded by the Purdue Research Foundation.

It should also be noted that once such a facility is in existence, it offers very high potential for use as an education tool, to provide short courses and continuing education modules

to logistics practitioners in the field. These courses could become a significant source of revenue, especially when we consider ongoing distance education activities through the CEE program as a potential distribution channel.

**Description of Deliverables**

This project will produce two deliverables. The first is a model for interaction and problem solving in the TDL domain, along with facilities to support this interaction. We will set up an environment for exploratory data analysis on e-Enterprise center equipment. This environment will consist of a database appropriate for transactional data, visualization and data analysis tools, and the glue to make them accessible to researchers on the project. Initially we will use already licensed software such as IBM's DB2 and Intelligent Miner, and freely available research tools including the DEVise visualization and data analysis tools from the University of Wisconsin – no software or equipment purchases are anticipated under this proposal. We plan this to be a permanent facility within the e-Enterprise center, with future support coming from external funding.

We will also deliver a simple, proof of concept prototype demonstrating the proposed environment and methodology, allowing us to illustrate how transactional data sets can be used to address issues of collaboration in the TDL sector. We initially plan to study the "fragmented temporal pools of capacity" problem previously described in the context of small trucking companies. This study, in additional to having potential impact by itself, will serve as a demonstration to attract real data and/or funding from industrial partners to support further development of this facility.

We will produce at least two proposals for further support of this work and the related facility. One will be for state funding (the Central Indiana Corporate Partnership and 21st century technology fund are planned targets.) We anticipate the second will be to a federal agency, although we will explore opportunities for corporate sponsored research as well.

**Investigator Credentials**

**Ananth Iyer** is Professor of operations management at the Krannert School of Management at Purdue University. He is the President of the Manufacturing and Service Operations

Management Society, the largest society within INFORMS. . Professor Iyer received his PhD in Industrial and Systems Engineering from Georgia Tech, his Masters from Syracuse University and his undergraduate degree in Mechanical Engineering from IIT Bombay. He was named a University Faculty Scholar at Purdue University in 1999. Prior to his current position, Professor Iyer was on the faculty at the Graduate School of Business at the University of Chicago. Professor Iyer chaired the committee to define the ecommerce option at the Krannert School. His research interests are in the area of logistics and supply chain management including understanding the effects of information sharing, incentives, contractual agreements etc in improving the performance of logistics systems. He has served as associate editor of *Operations Research, Manufacturing and Service Operations Management* and *IIE Transactions*. He has over 20 publications in *Operations Research, Management Science, Networks, European Journal of Operations Research, Information Processing Letters, Discrete Applied Mathematics, etc*. He has over 35 presentations in National and International Conferences as well as in various universities around the world. He has taught in executive programs in the U.S, China, Germany, Hungary, Spain and Argentina. He has worked as a consultant to food brokerage companies, catalog companies, cable television companies and transportation companies. He has also done pro bono work for the Chicago Public School System and the department of Streets and Sanitation in the city of Chicago.

**Reha Uzsoy** is a Professor in the School of Industrial Engineering and Director of the Laboratory for Extended Enterprises at Purdue University. He holds BS degrees in Industrial Engineering and Mathematics and an MS in Industrial Engineering from Bogazici University, Istanbul, Turkey. He received his Ph.D. in 1990 from the University of Florida and joined the faculty of Purdue University the same year. His teaching and research interests are in production planning and scheduling, and supply chain management. Before coming to the US he worked as a production engineer with Arcelik AS, a major appliance manufacturer in Istanbul, Turkey. He has also worked as a visiting researcher at Intel Corporation and IC Delco. His research has been supported by the National Science Foundation, Intel Corporation, Hitachi Semiconductor, Harris Corporation and General Motors. He was named Outstanding Young Industrial Engineer in Education by the Institute of Industrial Engineers in 1997 and a University Faculty Fellow by Purdue University in 2001, and has received both the A.A.B. Pritsker Award for excellence in

undergraduate teaching as well as the J. H. Greene Award for Outstanding Graduate Instructor. He is currently serving on the Editorial Boards on *IIE Transactions on Scheduling and Logistics, Journal of Manufacturing Systems, International Journal of Computer-Integrated Manufacturing* and *Journal of Scheduling*. He is the author of one book, more than fifty archival journal papers and numerous other technical publications.

**Chris Clifton** is an Associate Professor of Computer Science at Purdue. Prior to joining Purdue University in 2001, he held positions at Northwestern University and at the MITRE Corporation. He received his Ph.D. from Princeton University in 1991, and Bachelor's and Master's from the Massachusetts Institute of Technology in 1986. His research interests lie at the conjunction of database, data mining, and security. His research has been supported by the National Science Foundation and various branches of the Department of Defense. He currently serves on the Editorial Board of the *Journal of Knowledge and Information Systems*, and has over 30 articles in refereed journals and conferences.