Preserving Privacy in Supply Chain Management: a Challenge for Next Generation Data Mining[1]

Madhu Ahluwalia, Zhiyuan Chen, Aryya Gangopadhyay[2], Zhiling Guo
{madhu.is, zhchen, gangopad, zgou}@umbc.edu

Abstract

In this paper we identify a major area of research as a topic for next generation data mining. The research effort in the last decade on privacy preserving data mining has resulted in the development of numerous algorithms. However, most of the existing research has not been applied in any particular application context. Hence it is unclear whether the current algorithms are directly applicable in any particular problem context. In this paper we identify a significant application context that not only requires protection of privacy but also sophisticated data analysis. The area in question is supply chain management, arguably one of the most important research areas in production and operations management that has enormous practical relevance. We examine the area of supply chain management and identify research challenges and opportunities for privacy preserving data mining in the next generation.

1. New frontiers of privacy preserving data mining

The area of privacy preserving data mining (PPDM) started with the seminal paper by [1] and prompted numerous research efforts since then. Although many fundamental questions still remain unanswered, the area has matured to the point where it must establish its relevance to larger societal needs, including those of businesses and industries. Most of the existing methods may need to be extended or modified before being directly applicable to real-world settings. In this paper we describe one such area, supply chain management (SCM) that encompasses the entire business processes of all industries that deal with products involving multiple inter-related trading partners. Although data mining is not part of every day tools that are applied in SCM, statistical data analysis for demand forecasting is an essential part of production planning, inventory management, and order processing in most industries.

Supply chain management covers a multitude of tasks ranging from procurement of materials to transformation of these materials into intermediate and finished products and the distribution of these finished products to customers. The objective is to manage and control the material and information flow along the whole supply chain, so that the right products can be delivered in the right quantities at the right places at the right time at minimal cost. Demand forecasting is an important task in the management and optimization of supply chains that has a huge impact on a firm's profitability under uncertain business environment. Since the party closest to the market has most information, system-wide information asymmetry exists in the supply chain. A well-known phenomenon, known to researchers and practitioners in operations management for many years is the "bullwhip" effect [2], which refers to amplified demand fluctuation from downstream to upstream trading partners caused by multi-point forecasting at each echelon of a supply chain. The bullwhip effect has caused supply chains in the retail industry as a whole and in textile retail in particular, to lose billions of dollars every year in lost revenues and inventory cost. Lack of information sharing has been identified as one of the major reasons leading to SCM inefficiency.

Many organizations have realized that sharing information with other supply chain partners can lead to significant cost reduction. Collaborative planning, forecasting and replenishment (CPFR) is a relatively new approach aimed at achieving accurate demand forecasts and improving supply chain operations by sharing demand relevant information between trading partners in the supply chain. The key information includes point-of-sales (POS) data, future planned sales promotions, or inventory adjustments that would not have been known to the upstream partners if not shared. With the enhanced information visibility into the replenishment planning processes beyond the usual order cycle, demand forecast accuracy can be greatly improved. The reduction in forecast error across the supply chain improves operational efficiency among the supply chain partners and, therefore, yields mutual benefits.

Although conceptually attractive, a major challenge is the trading partners' unwillingness to share detailed information with the perception that other parties can unfairly exploit the information for their own benefits. Private information is normally viewed as a source of competitive advantage and is not freely shared among supply chain entities without a proper incentive mechanism. Due to firms' unwillingness to disclose proprietary demand information, credible information sharing is always viewed as a big obstacle in effective

[1] Research supported in part by NSF grant IIS-IPS 0713345 to Zhiyuan Chen and Aryya Gangopadhyay.

[2] Corresponding author. Authors are listed in alphabetical order. Authors' address: Department of Information Systems, University of Maryland Baltimore County (UMBC), 1000 Hilltop Circle, Baltimore, MD 21250.

supply chain management. Firms have recognized the need to hide sensitive information before sharing databases.

The rest of the paper is organized as follows. The next section introduces supply chain management and the problem of bullwhip effect, which motivates the need for privacy protection methods in supply chain management. In Section 3 we discuss the challenges and future research topics for applying privacy preserving data mining methods in SCM.

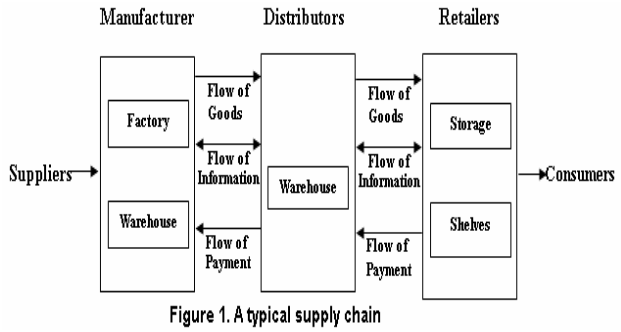

Figure 1. A typical supply chain

2. What are SCM and the "bullwhip effect"?

In a typical supply chain, there are five different types of entities: raw materials providers (i.e., suppliers), manufacturers, distributors, retailers, and customers. The raw material providers initiate the supply chain by drawing natural resources from the earth. Then, the manufacturers transform those resources into semi-finished or finished goods via conversion, manufacture, or assembly. The products then pass through necessary channels of distribution, often including warehousing. After some form of storage and delivery, the goods arrive at retail outlets. And the cycle ends with consumption and recycling by the consumer.

As shown in Figure 1, a traditional supply chain has three distinct dimensions [3] – the actual physical distribution of tangible ("hard") goods with inbound and outbound logistics systems, the exchange of currency or payment, and the exchange of information among various economic players. As raw material flows downstream from raw material suppliers through the supply chain to the manufacturers, it is transformed into more functional and integrated products with a higher economic value. Further downstream, it flows through distribution channels to retail outlets, and finally reaches the consumer. Information can flow from retail outlets to the trading partners upstream in the form of market forecasts and orders, and also from suppliers/manufacturers to the trading partners downstream in the form of order status and shipment information. These information flows have a direct impact on the production scheduling, inventory control, and delivery plans of individual members in the supply chain. In order to meet consumer demand, a large number of suppliers and manufacturers must work together to manage the flow of material and information. Without proper streamlining of the information and material flow in this highly complex supply chain, billions of dollars can be lost in the form of stockouts, defects, mark-downs, and inventory costs.

While the above sequence of business processes describes a supply chain, supply chain management refers to planning, design, and control of the flow of information and materials along the supply chain in order to meet customer requirements in an efficient manner. In traditional supply chain management, distributors play an important role in providing a shipment consolidation/integration function. Distributors collect orders from the retailers, fill the orders from their own warehouse inventory, and order products with the manufacturers. Since out-of-stock merchandise results in lost sales and possibly lost customers, distributors must be able to supply retail product demands quickly from inventory on hand. Thus, distributors have to maintain large inventories in warehouses as a buffer against demand uncertainty and possible product delivery delays by manufacturers. Accurate forecasts on both the retailers' orders and the end consumer market demand have an effect on the distributors' efficient inventory management.

Distributors usually adopt a periodic review inventory policy. When the inventory level is lower than a specific amount, distributors order products from manufacturers. Distributors place orders with manufacturers based on two important criteria: the retail demand and the wholesale price. Thus, distributors must forecast both future demand and future manufacturer pricing levels. Generally, distributors order products from manufactures in full truckload quantities to minimize shipping costs. There is also a processing cost for a purchase transaction. These factors contribute to orders in large batch sizes that do not reflect real demand. To take advantage of the trade promotions (i.e., wholesale price discounts) provided by the manufacturer during a short period of time, strategic distributors tend to order with a deviation from actual demand. The distorted demand information can be a problem to the manufacturers, as it leads to uneven production schedule and unnecessary inventory cost. This can cause one of the biggest problems in traditional supply chain management, termed the "bullwhip effect" [2], a phenomenon that creates fluctuation of order information and is amplified from downstream to upstream in the supply chain. Sharing POS data, exchange of inventory status information, order coordination, and simplified pricing scheme can help mitigate the bullwhip effect. However, it remains a challenging question as to why the downstream players in the supply chain would provide upstream partners with the necessary but possibly sensitive data.

3. How can PPDM be used in SCM?

Even though retailers can build a trusted relationship with their suppliers as in the case of Walmart, sharing proprietary information is generally not simple, and in fact can be a big risk if shared with competitors. Therefore, one useful strategy in business is to present information relevant for decision making without disclosing unnecessary details and at the same time ensuring that useful information is not lost. Privacy preserving data mining may hold the key to such a solution. However, there are several challenges in applying existing methods of privacy preserving data mining to SCM. We provide further details on some of the related issues below.

3.1 Integrated framework

One challenge of applying PPDM techniques in SCM is the inability to provide an integrated approach that works for many different analysis or mining methods. Typically more than one method is used in SCM to ensure accuracy of predicting future demand. Some of these methods are discussed in Section 3.3. In the literature on PPDM, there has been a rich body of work to address the privacy issue in distributed environments [4-15]. However, this work addresses the issue by sharing intermediate mining results to calculate mining functions securely over multiple sources. For example, in [16] the shared information is cluster models, in [8] the shared information is distances of each point to the cluster centroids, in [15] the shared information is Bayesian learning models in the context of Naive Bayesian learning, and in [10], it is the binary vectors used in the decision tree learning algorithm. Thus, each of these methods suits only a specific data mining task which may not be useful if a completely different mining or analysis method is needed. The utility and the need for an integrated approach is stated in [17] from the perspective of flexibility and better usefulness of privacy preserving data mining algorithms.

There have been some studies on integrated methods. For a non-distributed environment, a condensation approach has been proposed in [18]. This approach first generates size-k clusters and then regenerates data based on the properties of these clusters. However, multiparty computations are not considered in this work, and disclosure protection of the original data values may not be good enough for privacy because the regenerated data values are very close to the original ones. A transform-based approach for distance-based analysis or mining methods has been proposed by Liu et al. [19] and by us [20]. However, it is not clear how this approach can be applied to non distance-based methods. Therefore, there is still a need for research to find an integrated approach.

3.2 Data aggregation and partitioning

A typical supply chain includes a few manufacturers serving a few dozen distributors, feeding into hundreds of wholesalers, supplying to thousands of retailers [21]. This natural hierarchy along the supply chain dimension creates several data-related challenges. First, the data gets aggregated from downstream to upstream trading partners. Without a thorough analysis, it is unclear what effect this aggregation would have on the sensitivity of the data. On the surface, aggregation may provide some amount of privacy. However, it is well-known that there is no guarantee that privacy can be provided just by aggregating the data. So the question is whether it is possible to devise novel aggregation techniques that can provide privacy while at the same time allow accurate data analysis that would enable effective decision making by each trading partner. Another issue is data partitioning. The literature on privacy preserving data mining has dealt with horizontal and vertical partitioning as individual cases. In a supply chain, both horizontal and vertical partitions occur in tandem. All attributes for downstream trading partners may not be applicable or needed by upstream trading partners. It is unclear how the secure multi-party computation framework would work under these situations.

3.3 Statistical Analysis of Data

A typical grocery store needs to forecast about 5,000 different items. Stockout of a particular brand or size of a package can cause a loss of sales for both the retailer and the producer. Since variation between forecasting and the actual demand needs to be absorbed by extra capacity, inventory or rescheduling of orders, forecasting is an important input and an integrated part of a firm's operational decision making on process design, capacity planning, and inventory control. Therefore, accurate demand forecasting can improve an individual firm's operational efficiency. Collaborative planning, forecasting and replenishment (CPFR) is proposed to reduce the bullwhip effect and improve the overall supply chain performance.

In addition to the qualitative forecasting methods that rely on managerial judgment, there are two types of quantitative forecasting methods: time-series and causal forecasting. The commonly used methods for time-series forecasting include moving-average, weighted moving average, exponential smoothing, etc. One of the basic assumptions of all time-series methods is that demand can be decomposed into components such as average level, trend, seasonality, cycle, and error. Each of these items, except the random component, would be estimated from past sales data to develop an equation that is used to project forward into the future demand. Data decomposition and aggregation is a frequently used technique in time-series demand forecasting methods.

Although point-of-sales data is not too sensitive to be shared and is often used in time-series forecasting, demand and sales are not always the same thing. When the demand is constrained by capacity or other management policies, the forecasting of demand will not be the same as the forecasting of sales. Therefore, forecasting based on past sales data is not accurate enough to predict the future demand. This is especially true in today's hectic business environment.

In general, causal forecasting methods develop a cause-and-effect model between demand and other variables. For example, the demand for one brand of ice cream may be related to population, summer temperature, store promotion, prices and availability of other substitute brands. Data must be collected on these variables to determine the underlying relationship. One of the best-known causal methods is regression. Other forms of causal forecasting such as econometric models and simulation models are used as well, even though they are more complex and more costly to develop than regression models. The challenge, however, is that some data are extremely sensitive. Data on consumer age and consumption habits are directly related to consumer privacy concerns. Data on store promotion and inventory replenishment are related to the retailer's management strategy. For competitive reasons, sometimes data about competing brands are not freely disclosed. Therefore, the types of data and the range of information that can be acquired or shared are very limited due to the nature of intrinsic data sensitivity.

Sophisticated statistical methods are continuously being developed to improve the supply chain forecasting accuracy. However, the real difficulty is the availability of quality input data. PPDM can play a significant role in handling such data sharing challenges. Our goal is to protect privacy of the sensitive raw data by techniques of data transformation, so that the original data is hidden but the necessary information can be shared among supply chain partners to yield better forecasting outcome.

4. Conclusion

This paper discusses an important research area in production and operations management, namely, supply chain management as a potential application for methods on privacy preserving data mining. While many algorithms have been developed in the area of PPDM, it is unclear as to how the current algorithms may perform, if directly applied in this problem context. We have identified several research challenges that have to be overcome by extending and modifying current algorithms on privacy preserving data mining before they can be applied to solve problems in SCM. Supply chain management is relevant to all industries dealing with manufacturing and distribution of goods, and ties entire stakeholder populations from suppliers of raw materials to consumers. Researchers in SCM have suggested that lack of trust among trading partners is one of the biggest challenges in information sharing in SCM, which in turn can and does introduce enormous losses in supply chains across industries. Research in PPDM can alleviate the trust issue and thereby remove a major obstacle in supply chain integration. This is clearly a major area where research in privacy preserving data mining can contribute and make a big impact on business and society.

References

[1] R. Agrawal and R. Srikant, "Privacy preserving data mining," presented at 2000 ACM SIGMOD Conference on Management of Data, Dallas, TX, 2000.

[2] H. Lee, P. Padmanabhan, and S. Whang, "Information Distortion in a Supply Chain: The Bullwhip Effect," *Management Science*, vol. 43, pp. 546-558, 1997.

[3] M. Warkentin, R. Bapna, and V. Sugumaran, "The information dimension of emerging supply chain relationships in E-Commerce markets," *Journal of Electronic Commerce Research*, vol. 1, 2000.

[4] X. Lin, C. Clifton, and Y. Zhu, "Privacy preserving clustering with distributed EM Mixture modeling," *International journal of knowledge and Information Systems*, vol. 8, pp. 68-81, July 2005.

[5] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 1026-1037, September 2004.

[6] W. Du and M. J. Atallah, "Secure multi-party computation problems and their applications: a review and open problems," presented at 2001 Workshop on New Security Paradigms, Cloudcroft, NM, 2001.

[7] J. S. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," presented at 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, July 2002.

[8] J. S. Vaidya and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data," presented at 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D.C, August 2003.

[9] W. Du and Z. Zhan, "Building decision tree classifier on private data," presented at IEEE International Conference on Privacy, Security and Data Mining, Maebashi City, Japan, December 2002.

[10] C. Giannella, K. Liu, T. Olsen, and H. Kargupta, "Communication Efficient Construction of Decision Trees Over Heterogeneously Distributed Data," presented at Fourth IEEE International Conference on Data Mining, 2004.

[11] D. Caragea, A. Silvescu, and V. Honavar, "Decision Tree Induction from Distributed, Heterogeneous, Autonomous Data Sources," presented at Conference on Intelligent Systems Design and Applications, 2003.

[12] H. Kargupta and B. H. Park, "A fourier spectrum-based approach to represent decision trees for mining data streams in mobile environments," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 216-229, 2004.

[13] R. Wright and Z. Yang, "Privacy-preserving Bayesian network structure computation on distributed heterogeneous data," presented at 10th ACM SIGKDD Conference (SIGKDD'04), Seattle, WA, August 2004.

[14] D. Ma, K. Sivakumar, and H. Kargupta, "Privacy sensitive bayesian network parameter learning," presented at 4th IEEE International Conference on Data Mining (ICDM'04), Brighton, UK, November 2004.

[15] M. Kantarcioglu and J. Vaidya, "Privacy preserving naïve bayes classifier for horizontally partitioned data," presented at IEEE ICDM Workshop on Privacy Preserving Data Mining, Melbourne, FL, November 2003.

[16] S. Merugu and J. Ghosh, "Privacy-preserving distributed clustering using generative models," presented at 3rd IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL, November 2003.

[17] W. Du, C. Clifton, and M. J. Atallah, "Distributed Data Mining to Protect Information Privacy," presented at NSF Information and Data Management (IDM) Workshop, 2004.

[18] C. C. Aggarwal and P. S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," presented at 9th International Conference on Extending Database Technology, Heraklion, Crete, Greece, 2004.

[19] K. Liu, H. Kargupta, and J. Ryan, "Random Projection-based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," *IEEE Transactions on Knowledge and Data Engineering.*, vol. 18, pp. 92-106, January 2006.

[20] S. Mukherjee, Z. Chen, and A. Gangopadhyay, "A Privacy Preserving Technique for Euclidean Distance-Based Mining Algorithms Using Fourier-Related Transforms," *VLDB Journal*, vol. 15, pp. 293–315, 2006.

[21] C. B. Crespo-Marques A., and J.N.D. Gupta, "Operational and financial effectiveness of e-collaboration tools in supply chain integration," *European Journal of Operational Research*, vol. 159, pp. 348-363, 2004.