All That Glitters is Not Gold: Digging Beneath the Surface of Data Mining

Anthony Danna
Oscar H. Gandy, Jr.

ABSTRACT. This article develops a more comprehensive understanding of data mining by examining the application of this technology in the marketplace. In addition to exploring the technological issues that arise from the use of these applications, we address some of the social concerns that are too often ignored.

As more firms shift more of their business activities to the Web, increasingly more information about consumers and potential customers is being captured in Web server logs. Sophisticated analytic and data mining software tools enable firms to use the data contained in these logs to develop and implement a complex relationship management strategy. Although this new trend in marketing strategy is based on the old idea of relating to customers as individuals, customer relationship management actually rests on segmenting consumers into groups based on profiles developed through a firm's data mining activities. Individuals whose profiles suggest that they are likely to provide a high lifetime value to the firm are served content that will vary from that which is served to consumers with less attractive profiles.

Social costs may be imposed on society when objectively rational business decisions involving data mining and consumer profiles are made. The ensuing discussion examines the ways in which data mining and the use of consumer profiles may exclude classes of consumers from full participation in the marketplace, and may limit their access to information essential to their full participation as citizens in the public sphere. We suggest more ethically sensitive alternatives to the unfettered use of data mining.

KEY WORDS: analytics, customer relationship management, data mining, marketing discrimination, personalization, price discrimination, privacy, profiles, public sphere

Anthony Danna is a recent graduate of the masters program at the Annenberg School for Communication at the University of Pennsylvania.

Oscar H. Gandy, Jr. is the Herbert I. Schiller Term Professor at the Annenberg School. He is the author of The Panoptic Sort, Communication and Race, *and an engagement with the ethics of identification published in the* Notre Dame Journal of Law, Ethics & Public Policy.

Introduction

This article examines issues related to social policy that arise as the result of convergent developments in e-business technology and corporate marketing strategies. As more firms shift many of their business activities to the World Wide Web (the Web), increasingly more information about consumers and potential customers is being captured in Web server logs. Sophisticated analytic and data mining software tools enable firms to use the data contained in these logs to develop and implement a complex relationship management strategy. Although this new trend in marketing practice is based on the old idea of relating to customers as individuals, customer relationship management actually rests on segmenting consumers into groups based on profiles developed through a firm's data mining activities. Individuals whose profiles suggest that they are likely to provide a high lifetime value to the firm will be provided opportunities that will differ from those that are offered to consumers with less attractive profiles.

Although there are some observers who invite a careful assessment of the costs and benefits that data mining represents for the corporation, only very limited attention is being paid to the distribution of costs and benefits that we might

Journal of Business Ethics **40**: 373–386, 2002.

observe at other levels of society. Developing a more comprehensive understanding of the social impact of this marketing technology is the focus of this article.

We begin with an examination of the ways in which data mining technologies are applied in the market to support corporate marketing strategies. Technological and application-related issues are taken up before we introduce a discussion of the social concerns that emanate from the application these technologies in the public and private sectors. We conclude with several recommendations for mitigating the negative social impacts of data mining.

How personalization programs work

Data mining technology is employed in a variety of analytic and customer relationship management programs that are sold directly to firms or offered through an application service provider. In their promotional literature, the software companies that sell these programs emphasize the need that both business-to-business and business-to-consumer enterprises have to build better and more profitable relationships with their customers in a customer-centric economy.

Analytics

Analytic software allows marketers to comb through data collected from multiple customer touch points to find patterns that can be used to segment their customer base. Call center, product registration, and point-of-sale transaction generated data are typical of the off-line touch point data used in this type of analysis. The Web is another touch point that creates vast amounts of data that firms are including in their mining activities. Web-generated data includes information collected from forms, transactions, as well as from clickstream records. Clickstream data[1] allows for path analysis, shopping cart analysis,[2] the analysis of entry and exit points, and the analysis of search terms or key words entered by a visitor. Through the use of cookies, firms can add technographic information to their database that includes a user's Internet connection speed, software platform, and Internet service provider (ISP) address. Data from these off-line and online touch points can also be mined with third-party demographic and psychographic databases[3] when aggregated into a data warehouse. Data mining algorithms can be run in these warehouses to discover hidden patterns and trends that are used to create consumer profiles.

Data mining is increasingly being seen as an essential business process. Firms awash in data are desperately trying to capitalize on it. Over half of *Fortune's* top 1000 companies planned on using data mining technologies in 2001 to help determine their marketing strategy, a substantial increase since 1999 when under a quarter of these firms used data mining as a knowledge discovery technique (LeBeau, 2000). The following describes the primary methods of data mining used by firms for knowledge discovery.

Neural networks and decision trees. Artificial neural networks are designed to model human brain functioning through the use of mathematics. In order to be applied as a data mining technique, neural network processing elements must first be trained to discover patterns and relationships by using a sample of data. The network is tested against a second set of data to validate the predictive model it has generated. How well the network performs in predicting values in the validation set is used as an indicator of how well the network will predict outcomes with new data. Because neural network technology has the capability of learning, it does not require intensive programming instructions to sort through data. Neural networks have been employed in communication research to predict television extreme viewers and nonviewers (Paik and Marzban, 1995) and in the financial services industry to develop credit-scoring criteria and to predict bankruptcy.

Like neural networks, data mining through the use of decision tree algorithms discerns patterns in the data without being directed. According to Linoff, "decision trees work like a game of 20 Questions," by automatically segmenting data into groups based on the model generated when the algorithms were run on a sample of the

data (1998, p. 44). Decision tree models are commonly used to segment customers into "statistically significant" groups that are used as a point of reference to make predictions (Vaneko and Russo, 1999).

Market basket analysis and clustering. Both neural networks and decision trees require that one knows where to look in the data for patterns, as a sample of data is used as a training device. The use of market basket analysis and clustering techniques does not require any knowledge about relationships in the data; knowledge is discovered when these techniques are applied to the data. Market basket analysis tools sift through data to let retailers know what products are being purchased together. Clustering tools group records together based on similarity/dissimilarity scores applied to each data point in the individual record. Linoff notes that "clustering is typically one of the first techniques applied; the segments found in clusters often prove useful" (1998, p. 44).

Campaign management and personalization

Firms find the segments found in clusters prove to be most useful when they are integrated into a marketing strategy. Campaign management and personalization solutions incorporate or use elements of analytic programs to allow marketers to engage in a targeted one-to-one dialogue with their customers. For the most part, Web personalization occurs in one of three ways: manual decision rule systems, collaborative filtering systems, and observational personalization systems. These three categories are not mutually exclusive as programs can combine elements of each. Manual decision rule systems serve personalized content based on static user profiles. Static user profiles contain information collected about a consumer over the course of that consumer's relationship with the firm. The profiles are static in that they are not altered as a result of the consumer's Web activities. Collaborative filtering systems serve personalized content based on an analysis of information provided by the consumer via a Web interface. Collaborative filtering systems typically use information collected in a registration process for analysis. Observational personalization systems analyze clickstream data and dynamically serve personalized content based on that analysis (Mulvenna et al., 2000). This data is then fed to a recommendation engine where it is compared to profiles of previous visitors in order to provide the current user with content that is predicted to match that user's preferences. Users can be anonymous or identifiable in this process. The following scenarios illustrate how personalization systems that rely on analytics and data mining may be used in marketing applications.

Manual decision rule systems. Bob and Alice are both customers of Main Street Bank. Like most financial services companies, Main Street Bank is interested in reducing its churn rate. By applying its information about Bob to the predictive models it generated through its data mining activities, the bank is able to identify Bob as a customer who has a high lifetime value ranking but is at risk of leaving the bank for another financial services provider. Alice, on the other hand, fits the profile of a low lifetime value customer and is also assumed to be unlikely to leave the bank.

Since the bank is concerned with reducing churn, it offers Bob an interest bearing checking account and a reduced loan rate as an enticement to stay; these offers are Main Street's effort to strengthen its relationship with Bob. In communicating these changes to Bob, the bank emphasizes the value it places on having him as a customer. Because Alice falls into a less lucrative category of customer, she is not served the same offers as Bob. In fact, because Main Street Bank's profile of Alice suggests that she is unlikely to leave the bank, Alice is served with a notice that the fee she is assessed for using a teller inside the bank will be increasing. In communicating these changes to Alice, the bank emphasizes the value it places on having her as a customer and reminds her that she can continue to use the ATM at no additional charge. Thus we see that Bob, who is perceived to be more valuable to the bank is rewarded with lower prices for the services he uses, while Alice's fees are likely to rise.

Collaborative filtering systems. A new customer, Ted, visits Groove.com, an e-commerce site that sells CDs. The site's homepage encourages new customers to register before making a purchase in exchange for a discount on shipping. Ted is enticed by the shipping offer and chooses to register with Groove.com before making his purchase. He completes an online registration form where he is asked to rank his musical preferences and rate different artists. After he is finished registering he browses the site for music. As a registered user, a correlation engine that uses his registration data to predict what his preferences might be determines the content Ted is served as he navigates through the site. Ted's profile is updated each time he views information about an artist, or downloads a sample. In addition, each time he makes a purchase he provides the correlation engine with new data that can be used to narrow the range of options he will be presented with on his next visit.

Observational personalization systems. Carol is interested in purchasing a new computer and she visits TechStation.com, an electronics e-tailer. Carol is a first-time visitor to this site. After entering a few keywords to search the site and after browsing through several of the pages she selects the model she is interested in. Carol adds a printer to her virtual shopping cart and continues browsing. The observational personalization system used by the electronics store compares her point of entry to the site, the keywords she used in her initial search, her clickstream within the corporate site, and the contents of her shopping cart to the navigational patterns of existing customers already in firm's database. Through this comparison, the system fits Carol into the "young mother" profile that it developed by mining the Web navigation logs generated by previous visitors and existing customers. Accordingly, the recommendation engine offers Carol a discounted educational software package before she checks out.

Carol was, in fact, not a young mother, but a middle-aged divorcee. She purchased the computer and printer she was interested in, but did not find the time management software she actually wanted to buy. A bit frustrated, Carol leaves the site in search of the software she needs. At about the same time, Steve entered the site and selected the same computer and printer. Although he chose the same products as Carol, Steve did not receive the same offer for discounted educational software. He entered the site from a different portal than that used by Carol; he had a different clickstream pattern from hers, and he used different terms in his keyword search. Steve's navigational pattern resulted in his being assigned to a different profile. Steve fit best into the "college student" profile and as a result, he was offered a discount on a statistical software package. In fact, Steve is an English major. Like Carol, Steve's projected needs did not accurately match his real needs.

As these scenarios suggest, all systems will err to some degree when they attempt to predict individual interests and needs. The consequences that flow from the accumulation of such errors are at the base of the concerns we will discuss below.

Perceived benefits of personalization and campaign management

Customer relationship management

The software companies that market personalization products that use data mining techniques for knowledge discovery, speak to their potential clients in a language that make the benefits of these systems unmistakable. The promotional materials these firms use on their Web sites employ the language of customer relationship management. Market share and the development of market share by acquiring new customers was once the primary driver of marketing strategy. However, customer relationship management appears to be the philosophy that will drive marketing strategies in the 21st century. Customer relationship management focuses not on share of market, but on share of customer. Marketing strategists have been able to demonstrate that a firm's profitability can increase substantially by focusing marketing resources on increasing a

firm's share of its customers' business rather than increasing its number of customers (Peppers and Rogers, 1993).

One of the basic tenets behind customer relationship management is the *Pareto Principle*, the notion that eighty percent of any firm's profit is derived from twenty percent of its customers. Engaging in a dialogue with that twenty-percent in order to ascertain what their needs are and offering goods and services to meet those needs are said to be what customer relationship management is all about. Data mining technologies have allowed firms to discover and predict whom their most profitable customers will be by analyzing customer information aggregated from previously disparate databases. The Web has created a forum for firms to engage in a one-to-one dialogue with particular segments of their customer base in order to ascertain what the needs of those segments are.

To better understand how segmenting for the purposes of customer relationship management might be actualized, we analyzed the promotion materials posted on the Web sites of forty analytical, personalization, and e-commerce software vendors. The sample was generated from trade press articles on customer relationship management and from *Hoover's* online directory. Our analysis of the software companies very heterogeneous Web sites revealed five recurring themes that appear in their promotional material. Using similar buzzwords and phrases, these themes speak to the essential elements of customer relationship management and the ways in which e-businesses can use Web to make the strategy work to their advantage.

Ease of use. The technology that makes analytical and personalization software work can involve very complex processes including the use of algorithms, online analytical processing, and neural networks. This complexity can be intimidating for, and beyond the area of expertise of staff in the marketing department who are ultimately responsible for implementing these technologies. The software firms, for the most part, have designed their programs with this in mind. Interfaces have been designed specifically for use by marketing analysts. Most programs come with pre-installed reporting tools that can be easily customized. PrimeResponse summarizes this theme best by noting that its product minimizes "dependency on the IT department and puts power back into the hands of your marketing staff" (PrimeResponse, 2001).

General to specific needs customization. Dialoging with customers is how the "loop" in the marketing process is closed. A firm starts with some understanding of what its aggregate customers' needs are. This is usually the result of market research. The next stage in the loop involves developing a product or service to meet those needs. Communicating that product or service offering to customers follows that process. The loop is closed when a firm gets feedback from its customers and uses that feedback to refine its product or service offering. Feedback can come in the form of a direct answer to a question concerning needs or it can be indirect and processed in such a way that it is used to discover needs.

Analytic and personalization software programs allow firms to dialogue with their customers and refine their product or service offerings in a one-to-one fashion. By using analytical and personalization software, e-businesses can determine what an individual customer's needs are – based on the profile the customer fits into – and provide that customer with a product or service that meets those needs. Peppers and Rogers (1993) call this type of feedback collaboration. In their view, the firm and its customers collaborate to meet each other's needs.

360° view of customers. Integrating data from the Web and other customer touch points can give firms a more holistic view of their customers. This information is used to segment the customer base into communities of customers with similar characteristics (Peppers and Rogers, 1997). The more information a firm has about its individual customers, the easier it will be to create profiles and place individuals into them. As discussed above, these customer profiles, the product of analytic programs, are essential tools in the personalization process. Although the personalized

content a customer receives may appear to be unique, it is specific rather than unique because it is based on these profiles. None of the personalized content served to a customer on a Web site is truly unique to that individual. It is specific to the group that the customer is determined to belong to (Newell, 2000).

Internet time analysis. Analytic and personalization software allows firms to respond to their customers in "Web-time." The speed at which Web log data can be analyzed and compared to stored profiles to serve up personalized content gives customers a seamless experience as they move through a firm's site. Speed is also an important variable in dialoging with customers. If a customer does not respond to personalized content in the way predicted by the profile she is assigned to, new iterations of the calculations run by the recommendation engine can be run to instantaneously provide her with new content (Greenberg, 2001).

Measuring return on investment. More software companies emphasize measuring return on investment (ROI) than any of the four themes listed above. While ROI receives a lot of attention in the marketing of analytic and personalization software, calculating the ROI for e-business initiatives is complex and as a result there is little consensus as to how it should be done (Peppers and Rogers, 1997, pp. 384–387; Greenberg, 2001, pp. 265–266). Regardless of how a firm chooses to calculate ROI however, a customer's lifetime value measurement is always part of the equation (Newell, 2000, p. 58). According to Newell, a customer's lifetime value is "perfect for calculating ROI for CRM programs because everything aimed at strengthening the customer relationship has the objective of increasing the customer's profitability over time" (2000, p. 58). Regardless of how a firm mines its databases to segment its customers, an estimation of lifetime value is part of each profile created. The idea that some customers are worth more than others is the foundation for customer relationship management.

Market conditions and competition

The competition between software companies that market analytic and personalization software is fierce. In order to differentiate themselves in the marketplace, companies that offer analytic programs have partnered with companies that offer personalization programs to provide firms with the most sophisticated technology available to implement customer relationship management programs. However, a white paper published by the Patricia Seybold Group in 1999 predicted that these best-of-breed partnerships would dissolve in favor of integrated applications (Harvey, 1999). That prediction, to some extent, has come true (Gonsalves, 2001a; Gonsalves, 2001b). Software companies that specialize in Web analytics and personalization will likely suffer the fate of net.

As more bricks-and-mortar companies expand their Web presence and transform themselves into bricks-and-clicks e-businesses software firms like SPSS that can supply a complete portfolio of data mining, analytics, and Web analytics will become market leaders. With the growth in knowledge discovery and personalization, professional services firms have established consulting practices to help firms capitalize on this new technology.

Technological and application issues

Like any new technology, the software programs discussed herein are not error proof. First, applying data mining algorithms to data from disparate databases is not a simple task. Error is likely when data needs to be extracted from disparate databases, loaded into a data warehouse, and cleaned prior to being mined (Young, 2000). This is especially the case when data from a Web traffic log is integrated with other data sources (Tillett, 2000; Fink and Kobsa, 2000). Second, if there is no way to separate good data from bad data, erroneous data that finds its way into a data warehouse is just as likely to be subjected to data mining algorithms as is more accurate data. There are myriad sources for error in data collection, and in many cases it is virtually impos-

sible to parse out inaccurate data. Third, there is the case of missing data. Econometricians and statisticians have developed a sizable literature that addresses various methodologies for fitting models involving missing data; however, the more important questions in this area revolve around the sources of missing data. Are particular classes or populations of people more likely to have missing data associated with their records? If so, how do the statistical methodologies employed to fill in these missing data points affect the data mining-derived knowledge generated about these groups?

With substantial room for error, recommendation engines will have difficulty coming up with recommendations that their targets will find appropriate. An error in data cleaning, for example, could easily alter the information stored about a particular customer and ultimately affect how that customer will be segmented. What a recommendation engine determines is the "right" content for a consumer will not necessarily be considered "right" in the eyes of that consumer. Berry and Linoff note that predictions based on data mining activities are nearly always wrong "at the level of individual consumers" (2000, p. 20). They argue that the benefits derived from the small percentage of predictions that are "right" outweigh the costs of not having made any predictions at all. This small percentage can generate a notable increase in sales or click rates for a firm. When measured in an actionable outcome like click-through or purchase dollars these predictions can then be, on average, "right" for the firm. Using actionable outcomes as measurement tools, firms can gauge the appropriateness of a recommendation that is based on a prediction. If a recommendation generates a sale for instance, it is right for the firm. If it does not, it is wrong. This information is then added to the consumer's profile and is used to calculate new recommendations.

Given the room for error in data, it is unlikely that a recommendation based on a prediction will precisely capture the needs of an individual consumer. Data stored in a record will never be able to truly represent a complex autonomous individual. Although the personalization software companies claim that a 360° view of the customer is possible, such a view can never be complete. Transaction-generated data provides a historical snap-shot. Its predictions of the future are based on the past, and we know the past is often an unreliable guide to the future. Predictive models rarely take human serendipity into account, and it is virtually impossible to predict the circumstances that will shape an individual's choices in the future.

Social concerns

When presented to firms, the principles of customer relationship management and the software tools that allow it to be implemented in an e-business setting appear to be a rational means to a profitable end. Yet, there are always some social costs that are imposed on society when business decisions based on data mining and consumer profiles are made. The ensuing discussion will focus on the ways in which data mining activities and the use of consumer profiles systematically excludes classes of individuals from full participation in the marketplace and the public sphere.

We suggest that price and marketing discrimination result from the profiles generated by data mining practices. The sorting and allocation of information and opportunity that are the consequence of data mining can be thought of in terms of an invitation. For some categories of persons, price discrimination is an invitation to leave quietly. For many of these same persons, "Weblining" and marketing discrimination ensures that invitations are rarely if ever addressed to them. As we will discuss, these outcomes raise fundamental concerns about fairness or distributive justice (Hausman and McPherson, 1996; Hochschild, 1981; Roemer, 1996).

Price discrimination – an invitation for exclusion

Most economists would agree that price discrimination occurs when the same good or service is sold to different consumers at different prices. In a classic work on the subject, Stigler (1966) observes that discrimination necessitates

the separation of consumers into two or more classes whose valuations of the good or service differ. Stigler notes that this segmentation "requires the product sold to the various classes differ in time, place, or appearance to keep buyers from shifting" (1966, p. 209). In addition to price sensitivity, the time, place, or appearance of a product will determine a consumer's valuation of that product and assigns that consumer into a class. From our discussion of data mining practices, it is easy to see how these practices can be used to classify consumers based on an estimated or predicted valuation. Stigler defines price discrimination as "the sale of two or more similar goods at prices which are in different ratios to marginal cost" (1966, p. 209). Using this definition, either the price, the good or service, or a combination of both could vary across consumer classifications.

In order to sell similar goods or services at prices that differ in their ratio to marginal cost firms must have, according to Varian (1989), some degree of market power, the ability to sort consumers, and the ability to prevent arbitrage.[4] Market power can come in the form of monopoly or oligopoly, either within an industry or for a product or service itself. Sorting consumers by valuation can happen in several different ways. Sorting consumers into classes through data mining is the most sophisticated method of classifying consumers by their valuation; however, price discrimination will generate marketplace disparities even when consumers self-select their class. Arbitrage is not an issue for service providers like banks, but it is a major issue for firms that sell information products and other goods that can easily be resold.

Economists have formally defined three types of price discrimination. The most common form is third-degree price discrimination. In cases of third-degree price discrimination, firms exploit the differences in price sensitivity they have identified in the marketplace. Student or senior citizen discounts realized by self-selection are examples of third-degree price discrimination.

In second-degree price discrimination, firms are able to further exploit the differences in price sensitivity and extract more surpluses from consumers by versioning their product. Airlines are the most common practitioners of second-degree price discrimination. Although all seats on an airliner move passengers from point A to point B, airlines version their fundamentally equivalent products through fare restrictions. For example, Southwest Airlines lists seven different fares,[5] each with different restrictions for travel between Baltimore and Chicago. Southwest has identified different fare classes and passengers self-select their fare class according to their valuation of the restrictions. Incidentally, Southwest also practices a form of third-degree price discrimination by offering children, infant, and senior citizen fares. This particular form of price discrimination does not generally raise ethical concerns about fairness because those who receive these travel discounts are presumed to have limited resources, and might not otherwise be able to travel.

Windowing in the film industry provides another example of second-degree price discrimination where market power, consumer sorting, and arbitrage all come into play (Owen and Wildman, 1992). Windowing allows the producers of programming to sell different versions of the same product to consumers in order to extract maximum profit. Film distributors stagger the release of a film through different channels (theater, video/DVD, pay-per-view, cable, broadcast, syndication) in order to differentiate or version their product. A producer has market power in that it has a monopoly on the market for a particular film. The staggering of release through different channels invites consumers to sort themselves into categories defined by the distributors (theater patron, renter, etc.) based on their ability and willingness to pay. By staggering release and employing copyright protections, producers have the ability to prevent arbitrage. It is clear in this example, as it is more generally that sellers always benefit from price discrimination (Meurer, 2001, p. 91). The questions becomes one of determining which, if any groups of consumers benefit. Rawlsian egalitarians might express some concern about this particular form of price discrimination because those with limited income are least likely to be provided any group-specific discounts (Baker, 2002).

First-degree price discrimination is the most

sophisticated form of this economic phenomenon in that it requires firms to perfectly exploit the differences in price sensitivity between consumers. The seller charges the buyer the highest price the buyer is willing to pay for a good or a service. Shapiro and Varian note, "it is awfully hard to determine what is the maximum price someone will pay for your product or service" (1999, p. 39). Enter knowledge discovery through data mining. Suddenly this awfully hard process becomes easier for one-to-one marketers who have to make decisions about the differential pricing of products and services in order to increase the firm's share of the high lifetime value customers' business.

If those customers who have a predicted high lifetime value are the ones a firm needs to keep, then those with a predicted low lifetime value are the ones a firm needs to get rid of or otherwise convert to a more profitable status. Many firms come to the conclusion that low margin customers are not worth the effort necessary to turn them into high margin customers. The easiest thing to do is to entice those customers to leave (Newell, 2000, p. 42). This is often achieved through price discrimination.

Peppers and Rogers (1997) have recommended placing customers into a three-tier hierarchy, based on a calculation of potential value: Most Valuable Customers, Most Growable Customers, and Below-Zeros. According to Peppers and Rogers, Below-Zeros represent "the flip side of the Pareto Principle – the bottom 20 percent who yield 80 percent of losses, headaches, collection calls, etc." (1997, p. 416).

The financial services industry is skilled in the art of price discrimination. This skill is the result of data mining technologies that help to segment customers (Peppard, 2000). Rogers (2001) describes efforts by the Royal Bank of Canada to take a more customer-centric approach to management by tiering its customer base in order to better channel communication and services. First, the bank mined its databases and developed an algorithm to model the lifetime value of its customers and to estimate the "growability" of certain segments. With this data in hand, the bank set out to differentiate its offerings. According to Rogers, the bank "has nudged more than 60 percent of its customers [that were paying on a fee-for-service basis] into flat-fee packages" because customers with flat-fee packages tend to stay loyal to the bank (2001, p. 1).

The Royal Bank of Canada was not concerned about retaining the loyalty of the 40 percent of its fee-for-service customers it did not nudge. This is an example of what Peppers and Rogers (1997) call "firing" the customer; in this case the bank made certain that a segment of its customer base had a disincentive to stay. We would suggest that the formal definition of price discrimination Stigler (1966) offers is broad enough to account for the actions the Royal Bank of Canada took to "fire" its customers. In an e-business setting, enticing certain customers to stay and others to leave can be accomplished through personalized content. Personalized content can take many forms, including that of the price tag associated with a particular product or service.

We have seen how data mining lends itself to customer relationship management. With its emphasis on increasing the firm's share of predicted high lifetime value customers' business, customer relationship management necessarily lends itself to price discrimination. Price discrimination and the segmenting of consumers for the purposes of exclusion need not only take place under the rubric of customer relationship management.

In the emerging market for digital information products and services, price discrimination will become increasingly popular as technology allows for versioning and differential pricing. In their guidebook to survival in the networked economy, Shapiro and Varian (1999), emphasize the need to version content and price in the production of information goods in order to maximize profitability. Cohen (2000) contends that price discrimination of this type would seriously restrict access to high quality products, especially for low-income consumers who would be priced out of the market or have no choice but to settle for products of lesser quality.

Whereas price discrimination in information markets is often justified in terms of its potential for increasing access to the segments of the

population that could not otherwise afford to purchase information goods, the evidence seems to suggest that quite the opposite result occurs. Those with more substantial resources are actually provided discounts or subsidies in order deliver their attention to advertisers who value them more (Baker, 1994, pp 66–69).

Weblining and marketing discrimination – no invitation at all

Discriminatory pricing strategies are only one way firms can exclude certain classes of consumers from the marketplace. Profiles can also be used to determine if a class of goods and services that are offered to them in the first place. In an e-commerce setting, it is commonplace for consumers to receive differential access to goods and services as the result of collaborative filtering or observational personalization techniques. Consumers who fit into a particular profile may not be offered certain goods as readily as those who fit into other profiles. Discrimination in access and service based on a constructed profile has consequences for people in physical spaces like neighborhoods as well as in administrated spaces like Web sites.

Take the case of Kozmo.com, the Internet-based home delivery service that closed its doors in April 2001. Kozmo was accused of geographical redlining by residents in several of the cities in which it offered door-to-door delivery of entertainment and food products. Although it had distribution centers located in predominantly African-American neighborhoods in both Washington, D.C. and in New York City, Kozmo did not make its services available to the residents of these neighborhoods. Its executives claimed the company had simply made a rational business decision and used neighborhood Internet usage as the basis for defining its service area. A more extensive evaluation of Kosmo's business practices suggested that there had been a pattern of discrimination in each of the cities in which Kozmo.com operated (Zaret and Meeks, 2000).

Marcia Stepanek (2000) of Business Week coined the term "Weblining" to describe how banks segment and rank their customers to differentiate services and their associated fees. While Weblining can encompass price discrimination, it is a more general term used to describe cases where classes of consumers are excluded from the marketplace. Invitations are systematically withheld. Like its bricks-and-mortar world counterpart, redlining, Weblining involves the categorical discrimination of groups based on characteristics of their neighborhoods rather than on information about specific individuals (Hernandez et al., 2001).

In the financial services industry, consumers profiled above some risk criterion level are unlikely to learn about lending programs and other credit offers. Lambert (1999) describes this process through his examination of the ways in which traditional, direct marketing practices are used to deliver credit and loan financing offers to desirable borrowers. A decision about who receives information about lending programs is made in a context where risk "is no longer defined in terms of default, but as the failure to be significantly profitable" (Lambert, 1999, p. 2185). Although lenders cannot by law prohibit consumers who do not receive direct solicitation from submitting an application, it is highly unlikely that many such applications would be forthcoming.

The fact that the law is most concerned about such discrimination only when the victims are members of protected groups does not erase the fact that many consider the practices to be unfair because people are not being treated as individuals capable of making a rational choice in their own interest.

In the case of redlining, a particular irony emerges that helps to make this point. An individual who is denied credit because of the neighborhood in which she lives may be unable to succeed with a claim of discrimination under civil rights law because she is not a member of a protected group. She is a victim, not because of her race, but because of the race of the people that live in, and help to determine the profile of her neighborhood.[6]

The societal impact of unfair discrimination by commercial firms varies with the nature of the goods and services involved, as well as with the populations that are denied quality, and over-

supplied with second-rate goods and services. We have suggested that access to information may be particularly troublesome in that it may limit consumers' ability to make informed choices in the marketplace, or to participate effectively within the public sphere.

Information in the public sphere

The Internet has often been hailed as an enhancement to the democratic public sphere. By its very nature, the Internet is thought to expand access to a broad range of voices and perspectives (Sparks, 2001). However, when it is used as a tool to segment and divide it can have a deleterious impact on the democratic process. Segmenting consumers for the purposes of delivering policy-related information creates and exacerbates inequalities that can distort public discussion and debate (Gandy, 2001). If information about public policies is more accessible to one class of citizens than to others, then the quality of the dialogue between and among the governed will most certainly be affected. The Web offers opportunities to lower barriers to access and engagement; segmentation only creates new barriers.

Sunstein (2001) makes the case that filtering on the Web erodes opportunities to have shared experiences and to lessen the likelihood that we will be exposed to viewpoints and information that we may not seek out on our own. These, he argues are essential to meaningful deliberation and to the functioning of a democratic society. Although he briefly discusses collaborative filtering techniques, Sunstein's argument rests primarily on the assumption that filtering decisions are made by individuals out of their own self interest. However, when a firm uses data mining techniques to create the profiles that are then used to serve filtered information the effects Sunstein describes become considerably more problematic. A profit-seeking firm will filter the information it supplies in its own economic interest, especially where those interests are tied to the interests of advertisers and investors (Baker, 2002). This filtering will not necessarily serve the private or collective interests of the individuals who are thereby informed, and it is quite likely to work against the interests of those consumers who have been excluded from the flow of information.

We believe the seriousness of this problem is amplified when it is the government that discriminates in the supply of information.

In his State of the State address in January 2001, California Governor Gray Davis announced the launch of California's new e-government portal, the "My California Homepage." BroadVision provides the site with personalization technology while Broadbase Software is providing the analytical tools necessary to evaluate visitor data and build profiles. California is not alone in implementing technology designed for the private sector in the delivery of governmental services and information through the Web. Delaware's Web portal incorporates content management solutions from Eprise Corporation (2001), who stresses the importance of using profiles to customize the delivery of Web content. Profiling and segmentation can result in some content being made extremely difficult for the average user to find. Indeed, in some cases, this information may have been placed "off limits" to particular classes of citizens.

For example, increased concern about monitoring and managing access to Web-based government information has developed following the events of September 11, 2001. Limitations on access to formerly public information are likely to be based on characterizations of users developed through data mining techniques (How Sept. 11 changed America, 2002; Gerstein, 2001).

As funding levels for intelligence agencies in the United States and abroad are increased in support of a new war against terrorism, it seems likely that many of these dollars will flow to business providing analytics software, data mining, and data warehousing resources (Gomes, 2002). Developments made in response to government contracts are likely to enhance the capability of data mining, data warehousing, and business analytics systems currently being developed for the commercial market. In addition, as the cost of existing systems drops, the use of these technologies will spread beyond the *Fortune* 1000 to commercial enterprises of varying size and

commitment to what we have come to recognize as fair information practices (OECD, 1980). As the use of these technologies become more ubiquitous the likelihood of their affecting commercial and civic life in the ways we have discussed will become even greater. Smaller firms are more likely to fall under the radar screen of watchdogs and are less likely to belong to the industry groups that have been assigned regulatory duties in the place of government.

Conclusions and recommendations

Decision makers should not discount the costs imposed on society when data mining and consumer profiles are used to identify and segment individuals into groups on the basis of estimates of value. Although progress toward establishing a reasonable expectation of privacy in transaction-generated information has been stalled in response to the terrorist threat, early attempts to formulate a policy on consumer profiling that would win the support of the business community have emphasized the importance of "notice and choice."

At the very least, consumers should be informed of the ways in which information about them will be used to determine the opportunities, prices, and levels of service they can expect to enjoy in their future relations with a firm. We note that a reliable test of the ethical status of any business practice is the extent to which it can be exposed to the light of public review. We might understand this test as an application of the Kantian standard of "universal acceptability" or the "Golden Rule" which admonishes us not to do anything to another than we would not have them do to us (Spinello, 1997, p. 37).

Our examination of discussions of data mining and segmentation techniques within the trade press reveals a broad awareness among the users of these techniques, that the public is concerned, indeed often outraged when they discover the ways in which they are graded, sorted, and excluded from opportunities that others enjoy (Berry, 1999). Because they have ignored this basic principal of mutual respect, many firms have been compelled to issue public apologies when the discriminatory nature of their routine business practices have been revealed in the press. Others, like DoubleClick have seen their stock fall out of favor with investors when discriminatory practices have been brought to light.

There are some signs that organizations whose very life depends upon proprietary technology for rating and ranking consumers recognize the importance of informing consumers about the ways in which their life chances are determined in the marketplace. Fair Isaac, the leader in credit scoring, has recently developed a commercial product that would inform consumers about the components of their credit scores, and how they might be improved (Simon, 2002). Of course, we are not suggesting that the problems of market discrimination that we have described can be overcome by selling access to information about the means by which such discrimination is accomplished.

We are recommending that every decision maker who bears any responsibility for implementing a marketing strategy based on data mining consider more than its impact on the bottom line. We are especially hopeful that decision makers would consider the Rawlsian principles of special regard for those who are least advantaged, rather then being guided, as they seem most often to be, by a utilitarian calculus that is blind to distribution (Hausman and McPherson, 1996; Roemer, 1996).

A Rawlsian perspective on social justice, guided by the bright light of publicity, would go a long way toward revealing the true value in the base metals that data mining has uncovered so far.

Notes

[1] Clickstream data represents the "footpath" a Web site visitor creates while navigating through a site. Data points are generated when site visitors click through the site, following links. For the typical e-business, clickstream data can grow to several terabytes in size. To put the size of these databases into perspective, it should be noted that two terabytes of data would be roughly equivalent to the amount of data stored in an academic research library. In addition to

being mined to generate information about customers, clickstream data is also analyzed to evaluate individual page traffic levels, gauge the effectiveness of site design, and evaluate the popularity of content.

[2] Shopping cart analysis uses clickstream data generated by e-commerce site visitors to investigate, among other things, the point in the visitors' footpaths where purchases are made and to evaluate the point at which shopping carts are abandoned by visitors.

[3] Psychographic data includes information about individual's attitudes, behaviors, and beliefs.

[4] Arbitrage is the term used to describe the resale of a product in a market where price discrimination has occurred. Arbitrage occurs when a consumer charged a lower price for a product resells that product to a consumer who does not have access to the product at that price.

[5] See http://www.southwest.com. Accessed on 10/03/01.

[6] Reference is made to the case of *Cherry v Amoco Oil Co.*, 490 F Supp 1026 (ND Ga 1980) in which a White woman who lived in a predominately Black neighborhood was denied a gasoline credit card. The zip-code was one of several factors included in the rating system used by Amoco. Her zip code received the lowest of five ratings assigned by the company, but she was unable to demonstrate that racial animus was the basis for the denial of credit.

References

Baker, C. E.: 1994, *Advertising and a Democratic Press* (Princeton University Press, Princeton).

Baker, C. E.: 2002, *Media, Markets, and Democracy* (Cambridge University Press, New York).

Berry, M. J. A and G. Linoff: 2000, *Mastering Data Mining* (Wiley, New York).

Berry, M.: 1999, 'The Privacy Backlash', *Intelligent Enterprise* (October 26), p. 20.

Cohen, J. E.: 2000, 'Copyright and the Perfect Curve', *Vanderbilt Law Review* **56**(6), 1799–1819.

Eprise Corporation: 2001, 'Customizing Web Information Delivery for Diverse Audiences' [On-line]. Available: http://www.eprise.com.

Fink, J. and A. Kobsa: 2000, 'A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web', *User Modeling and User-Adapted Interaction* **10**, 209–249.

Gandy, O.: 2001, 'Dividing Practices: Segmentation and Targeting in the Emerging Public Sphere', in W. Bennett and R. Entman (eds.), *Mediated Politics: Communication in the Future of Democracy* (Cambridge University Press, New York), pp. 141–159.

Gerstein, J.: 'Online Secrets. Internet Could Reveal Sensitive Information to Enemies', ABCNews.com [On-line]. Available: http://abcnews.go.com/sectio...ngAmerica/WTC_011015_InternetSecrets.html.

Gomes, L.: 'Siebel Hopes Government Will Choose its Software for the War on Terrorism', Wall Street Journal Online [On-line]. Available: http://online.wsy.com/article_print/0,4287,SB101942665887847 9920,00.htm.

Gonsalves, A: 2001a, 'Blue Martini Readies New CRM, Commerce Apps', TechWeb [On-line]. Available: http://www.techweb.com/wire/story/TWB20010209S0017.

Gonsalves, A: 2001b, 'Personify Readies Customer Analysis Product', TechWeb [On-line]. Available: http://www.techweb.com/wire/story/TWB20010130S0018.

Greenberg, P.: 2001, *CRM at the Speed of Light: Capturing and Keeping Customers in Internet Real Time* (Osborne/McGraw-Hill, Berkeley, CA).

Harvey, L.: 1999, 'E.piphany E.4: Comprehensive Advanced Customer Analysis Software for E-tailers', Patricia Seybold Group Information Assets, [On-line]. Available: http://www.epiphany.com.

Hausman, D. and M. McPherson: 1996, *Economic Analysis and Moral Philosophy* (Cambridge University Press, New York).

Hernandez, G., K. Eddy and J. Muchmore: 2001: 'Insurance Weblining and Unfair Discrimination in Cyberspace' *SMU Law Review* **54**, 1953–1972.

Hochschild, J.: 1981, *What's Fair?* (Harvard University Press, Cambridge, MA).

'How Sept. 11 Changed America, and What it Costs our Liberty', *The Wall Street Journal*, Online [On-line]. Available: http://online.wsj.com/article_print/0,4287,SB101555545794691712 0,00.htm.

Lambert, T.: 1999, 'Fair Marketing: Challenging Pre-Application Lending Practices', *Georgetown Law Journal* **87**, 2181–2224.

LeBeau, C.: 2000, 'Mountains to Mine', *American Demographics* **22**(8), 40.

Linoff, G.: 1998, 'Which Way to the Mine?', *AS/400 Systems Management* **26**(1), 42–44.

Meurer, M.: 2001, 'Copyright Law and Price Discrimination', *Cordozo Law Review* **23**, 55–148.

Mulvenna, M. D., S. S. Anand and A. G. Buchner: 2000, 'Personalization on the Net Using Web

Mining', *Communications of the ACM* **43**(8), 122–125.

Newell, F.: 2000, *loyalty.com: Customer Relationship Management in the New Era of Internet Marketing* (McGraw-Hill, New York).

Organization for Economic Cooperation and Development: 1980, 'Recommendations Concerning and Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data', in M. Rotenberg (ed.), The *Privacy Law Sourcebook 2000* (Electronic Privacy Information Center, Washington, DC), pp. 236–263.

Owen, B. M. and S. S. Wildman: 1992, *Video Economics* (Harvard University Press, Cambridge, MA).

Paik, H. and C. Marzban: 1995, 'Predicting Television Extreme Viewers and Nonviewers: A Neural Network Analysis', *Human Communication Research* **22**(2), 284–306.

Peppard, J.: 2000, 'Customer Relationship Management (CRM) in Financial Services', *European Management Journal* **18**, 312–327.

Peppers, D. and M. Rogers: 1993, *Enterprise One to One: Tools for Competing in the Interactive Age* (Currency, New York).

Peppers, D and M. Rogers: 1997, *The 1:1 Future: Building Relationships One Customer at a Time* (Currency, New York).

PrimeResponse: 2001, 'Solutions: E-tail & Retail', [On-line]. Available: http://www.primeresponse.com/solutions/retail.html.

Roemer, J.: 1996, *Theories of Distributive Justice* (Harvard University Press, Cambridge, MA).

Rogers, M.: 2001, 'CRM Dividends for Royal Bank of Canada', Inside 1to1, [On-line]. Available: http://www.marketing1to1.com.

Shapiro, C. and H. R. Varian: 1999, *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business School Press, Boston).

Simon, R.: 2002, 'Fair Isaac Plans Credit-score Help, But Watchdog Groups See Conflict', The Wall Street Journal Online [On-line]. Available: http://online.wsj.com/article_print/0,4287,SB1016487907757224760,00.htm.

Sparks, C.: 'The Internet and the Global Public Sphere', in W. Bennett and R. Entman (eds.), *Mediated Politics: Communication in the Future of Democracy* (Cambridge University Press, New York), pp. 75–95.

Spinello, R.: 1997, *Case Studies in Information and Computer Ethics* (Prentice Hall, NJ).

Stepanek, M.: 2000, April 3, 'Weblining', *Business Week*, pp. EB26–EB34.

Stigler, G.J.: 1966, *The Theory of Price* (Macmillan, New York).

Sunstein, C. R.: 2001, *Republic.com* (Princeton University Press, Princeton, NJ).

Tillett, S.: 2001, 'One-stop User Data Shop', InternetWeek, [On-line]. Available: http://www.internetwk.com/story/INW20010220S0003.

Vaneko, J. J. and A. W. Russo: 1999, 'Data Mining and Modeling as a Marketing Activity', *Direct Marketing* **62**(5), 52–55.

Varian, H. R.: 1989, 'Price Discrimination', in R. Schmalensee and R. D. Willig (eds.), *Handbook of Industrial Organization, Volume I* (Elsevier Science Publishing, Amsterdam), pp. 597–654.

Young, D.: 2000, 'The Information Store: CRM is Thrusting Data Mining Back to the Future', *Wireless Review* **17**(18), 42–50.

Zaret, E. and B. N. Meeks: 2000, April 4, 'Kozmo's Digital Dividing Lines', MSNBC [On-line]. Available: http://www.msnbc.com/news/373212.asp.

Annenberg School for Communication,
University of Pennsylvania,
3620 Walnut Street,
Philadelphia, PA 19104-6220,
U.S.A.
E-mail: ogandy@asc.upenn.edu