Original Article

Inequality in the utility of customer data: Implications for data management and usage

Received (in revised form): 6th January 2010

Adir Even

is a faculty member at the department of Industrial Engineering and Management at Ben-Gurion University of the Negev, Israel, since 2008. Adir, who received his doctoral degree from Boston University School of Management, explores the contribution of data resources to value-gain and profitability from both theoretical and practical perspectives, and studies implications for system design, data warehousing, business intelligence and data quality management. His research has been published in journals such as *IEEE/TDKE*, *CACM*, *CAIS*, the *DATA-BASE* and *DSS*.

G. Shankaranarayanan

is Faculty in the TOIM division of Babson College. His research interests cover modeling and managing data and metadata, managing data quality for decision support, and managing data using economic perspectives. He serves as the Area Editor of the *International Journal of Information Quality* and as an Associate Editor of the *ACM Journal for Data and Information Quality*.

Paul D. Berger

is a visiting scholar and professor of Marketing at Bentley University, where he is also the director of the Master of Science in Marketing Analytics program. His most recent (co-authored) book has just been published: *Internet Marketing: Reaching Customers Anytime, Anywhere, Any Platform*. In addition to the current text, he has published several previous texts, including *Experimental Design: with Applications in Management, Engineering, and the Sciences*. He has published over 150 refereed journal articles and conference proceedings, in such journals as *Management Science*, the *Journal of Marketing Research*, the *Journal of Interactive Marketing* and the *Harvard Business Review*, and is on the editorial boards of several journals, including the *Journal of Interactive Marketing*. He is an active consultant and won the Metcalf Award, a university-wide award for teaching excellence, and the John R. Russell award for excellence in executive teaching.

ABSTRACT Customer data has long been recognized as a critical resource for supporting marketing efforts. With the rapidly increasing investments in the acquisition and the administration of customer data, the concern about cost-effective management of this important resource is growing. In this study, we explore the concept of inequality in the utility of data – the extent to which records in large data sets differ in their business-value contribution. The Lorenz curve and the associated Gini index, powerful tools for analyzing inequality, have been applied to a variety of social and economic disciplines. We adapt these tools for modeling and measuring inequality in the context of managing large customer data sets, and use them for analyzing utility/cost trade-offs and optimizing data management decisions. We demonstrate the application of these tools in the customer relationship management context of managing alumni relations, using large data samples from a real-world system. The results support our argument that understanding and measuring inequality in the utility of data may have important implications for better usage and administration of customer data.

Journal of Database Marketing & Customer Strategy Management (2010) 17, 19–35. doi:10.1057/dbm.2010.1; published online 15 February 2010

Keywords: inequality; gini index; CRM; database marketing; data management

Correspondence:

Paul D. Berger
McCallum Graduate School
of Business, Bentley
University, 175 Forest St,
351 Morison, Waltham
MA 02452, USA
E-mail: pberger@bentley.edu

INTRODUCTION

The importance of collecting and managing customer data (for example, demographics, preferences, purchase transactions, contacts) is well recognized in the database/interactive marketing arena – very little can be done without the comprehensive support of a data resource. With the growing importance and the rapidly increasing volumes of customer data, organizations are required to invest in advanced technologies for processing, storing and delivering it. The large investments in customer data raise a question – do the benefits gained from the usage of customer data justify the data collection and administration costs? Are customer data resources being managed in an economically optimal manner?

This study is motivated by the need to consider economic aspects toward cost-effective management and utilization of customer data. Thus far, research has addressed data management primarily from a stand point of technical requirements (for example, storage and processing capacities, data management technologies) and functional requirements (for example, how to take advantage of data for decision support, what data is needed for the decisions that need to be made). In the database marketing field, various studies deal with how to best obtain customer data and what to do with such data when it is available. However, economic aspects of managing customer data, such as the costs associated and the benefits gained, have not been sufficiently studied.

As a contribution to that end, we examine data utility – a measure for the value gained by using data resources. Utility can be attributed, for example, to the use of data for enabling certain business processes, the improvement of decision outcomes or to the receipt of income from renting/selling data to other firms. We specifically address the issue of inequality in the utility of customer data resources,

arguing that ‘not all data are created equal’, as some subsets of customer data may be more valuable than others. In this study, we introduce quantitative tools for modeling and measuring the inequality in the utility of data – the extent to which different records in a data set differ in their utility contribution. These tools are adaptations of Lorenz’s curve and the Gini index – commonly used statistical tools for assessing inequality in large populations. We suggest that understanding the inequality within customer data resources and the associated utility/cost trade-offs have important implications for data management. From the data consumers’ viewpoint – it can inform better usage of data and help focus marketing efforts on the more profitable customer segments. From a data management viewpoint, it can impact the design of data environments, direct data acquisition and retention policies, and help prioritize data quality management efforts.

We demonstrate the concepts of measuring inequality and assessing utility/cost trade-offs with a large sample of a real-world data resource used for managing alumni relations. We show that the magnitude of utility inequality in this data resource is high, link it to current data acquisition and maintenance policies, and show how a utility-cost analysis can be used to evaluate data quality management alternatives and set an optimal policy. Although this study focuses on a specific context of customer data management, we suggest that utility inequality, its implications for data management, the statistical tools for assessing it and the evaluation methodology that we demonstrate are applicable in other contexts of customer data management.

In the reminder of this article, we first present the background and conceptual foundations that have influenced our thinking. We then present tools for assessing utility inequality and the associated utility-cost trade-offs in large data sets, with the goal

of optimizing data usage and administration decisions. We demonstrate an application of these tools for the aforementioned alumni data, showing that within this business context inequality assessment may have major implications. To conclude, we summarize the key contributions of our study, highlight its limitations and propose directions for future research.

BACKGROUND

Data repositories, along with the information systems (IS) utilizing them, have long been recognized as critical organizational resources. Recent years have witnessed a major transition toward extended usage of data resources for business analysis, performance measurement and managerial decision support. Davenport¹ provides examples of firms that have gained strong competitive advantage by investing in the development of data analysis capabilities and data-driven analytics. This transition toward data-driven management is well supported by the rapid progress in the capacity and the performance of information and communication technologies (ICT) for utilizing large data resources.

The efficiency of and the benefits gained from CRM and other IS environments that support database/interactive marketing depend on the data resources – customer profiles, transaction history (for example, purchases, donations), past contact efforts and promotion activities. Customer data supports critical marketing tasks, such as customer segmentation, predicting consumption, managing promotions and delivering marketing materials.² It underlies popular marketing techniques such as the Recency, Frequency and Monetary (RFM) analysis for categorizing customers,³ estimating Customer Lifetime Value (CLV)^{4–6} and assessing Customer Equity.^{7,8}

Business firms are profit-maximizing entities and hence, we view the maximization of economic benefits as an

important goal of data management, particularly in the context of managing customer data. Improving economic outcomes involves increasing the *utility*, the business benefits, gained from data usage, as well as reducing the *cost* involved in implementing and managing data environments (for example, ICT investments, data acquisition and maintenance costs). Some data management decisions may introduce significant economic trade-offs.^{9,10} For example, increasing data volume and richness, improving data quality and investing in advanced software platforms may improve utility, but involve higher costs.

We link economic trade-offs to the magnitude of *utility inequality* – whether utility contribution is the same for all records in a data set, or is relatively concentrated in a relatively small subset. The concept of assessing inequality of attributes within large populations has been examined in the contexts of income, property ownership, education and goods manufacturing, to name a few (for example,^{11,12}). It has also been used to direct data mining – automated exploration and analysis of large data sets.¹³ Beteille¹⁴ proposes a useful distinction between two aspects of inequality – the *relational* aspect and the *distributional* aspect. The sociologist is concerned mostly with relational inequality. Inequalities are seen as being built into the social structure in the form of relations of superordination and subordination, that is, the patterns of rights and obligations. The economist is concerned more with the distributional aspect of inequality, viewing inequality in the distribution of an outcome indicator such as wealth, income, health or educational status.¹⁵ We adopt more of the economist perspective for studying inequality in customer data. Given the large investments in this resource, one should be able to economically justify the expenses. We believe that understanding inequality

characteristics that exist within customer data in an organization can significantly change the manner in which data is managed today.

For modeling the distribution of utility and analyzing the magnitude of inequality, we adapt the Lorenz curve,¹⁶ and the Gini index¹⁷ – long-standing statistical tools used to analyze social and economic inequality in large populations. The *Lorenz curve* provides a mathematical formulation and a visual representation of the equality (or inequality) of distribution of a particular measure (for example, age, income, wealth) within a population (www.en.wikipedia.org/wiki/Lorenz_curve). The Gini index (also the Gini *number* or *coefficient*), which is derived from the Lorenz curve, is a [0,1] measure for the magnitude of inequality (www.en.wikipedia.org/wiki/Gini_coefficient). The higher the Gini index, the greater the inequality within the distribution represented by the Lorenz curve. A Gini index of 0 indicates a distribution in which the value of all items are identical, while a Gini index that approaches 1 indicates a distribution in which the value is concentrated within a small fraction of the population. The quantitative tools that we develop in the following section for modeling and measuring inequality in the utility of data are adaptations of the Lorenz curve and the Gini index. As our definitions are slightly different than those that are common in today's statistical literature (for example,^{12,13}) we preferred to use different terms for the curve and for the corresponding index.

INEQUALITY IN UTILITY OF DATA

The utility of IS and resources reflects their current and/or potential business-value contribution – the magnitude of improvement in business performance, decision outcomes and/or the information consumer's willingness to pay.¹⁸ Customer

databases are among the most important information resources in an organization today, being an essential input for marketing and CRM decisions. Given possible variability in the importance of items within a customer database (for example, tables, records and attributes), will the overall utility best be derived from the entire data resource, or only on a small subset of it? The response to this question represents the magnitude of *inequality in utility* within a customer database. To understand this notion of inequality in the utility of customer databases, we first illustrate it with a simple example, and then develop it further into a more general framework.

Utility attribution and understanding – an illustrative example

Our study addresses customer databases consisting of tabular data sets. The table is a common data storage model in databases, consisting of multiple records, each with the same set of attributes – for example, a list of customer profiles, in which the same demographic/income attributes are collected for each. While records in a tabular data set are similar in attribute structure, they vary in content. This content variability may differentiate the relative importance of records to data consumers and users; hence, their associated value contribution will vary as well.

The assessment methodology presented here is based on attributing *utility*, a numeric measure that reflects the relative importance and value contribution of a record from a business/usage perspective, to all data records. Utility estimation and attribution largely depend on the usage context. In many contexts, utility can reflect monetary value (for example, associated revenue potential). As explained later, assessing utility/cost trade-offs requires measuring both along the same monetary scale. However, the tools for modeling and

assessing inequality described in this study do not depend on the utility units. Several attribution methods (essentially, marketing metrics), reflecting relative importance and value, have been discussed in the literature and may be adapted for utility estimation in the context of database marketing; we already noted CLV (for example,^{3,4}) and RFM (for example,²) analysis, as examples. As the same data resource can be used for different purposes, it can be attributed with multiple different utility measures.¹⁹ For brevity, in this study we use a single-utility measure attribution, representing one usage or an aggregation of multiple usages.

We consider a tabular data set with N records (indexed by $[n]$), and assign each record a non-negative utility measure ($u_n \geq 0$), which reflects the relative importance of record $[n]$ in the evaluated usage. We assume no interaction effects between records and, hence, additivity, to arrive at the overall data set utility: $u^D = \sum_n u_n$. The data set utility u^D is maximum when the entire data set is available and may be reduced to some extent if some records are missing or defective. A simple utility allocation may assign an identical value per record (i.e., a constant $u_n = u^D/N$). While easy to compute, this 'naïve' allocation rarely reflects real-world data use, as records differ in importance and utility contribution.

To illustrate the concepts of utility attribution and inequality, we consider a sample (shown in Table 1) from a data set

that lists customer profiles. By analyzing certain demographic attributes and past purchases, the CLV can be estimated for each customer. This estimation is important for identifying a subset of the profiles for certain promotional campaigns, as well as identifying customers with a significantly higher purchase potential. In this example, we use the CLV estimation as a proxy for the utility of each profile. To what extent do profile records in this data set differ in utility? We sort them in descending utility order and calculate the cumulative sum. Figure 1(a) shows the cumulative utility vs the number of records. As the number of records reaches the maximum (here, 10), the cumulative utility reaches the total sum (here, 200). Notably, the magnitude of inequality appears to be high – utility ranges between 1 and 50. The first record (A) accounts for 50 out of 200 units while the three records with the lowest utility (C, J and E) account for 14 units, combined.

We now express utility as a ratio instead of absolute units. We rescale the curve by dividing the number of records (the horizontal axis) by the total (here, 10), and the cumulative utility (the vertical axis) by the total (here, 200). The resulting *cumulative utility curve* (Figure 1(b)) does not depend on the absolute utility or the unit for measuring it, but rather on the *relative* allocation. In this example, ~80 per cent of the utility is obtained from only 50 per cent of the records. This ratio highlights that in certain data sets, a relatively high

Table 1: The utility of customer profiles

#	Profile ID	Utility (CLV)	Cumulative utility	Relative utility	Cumulative relative utility
1	A	50	50	0.250	0.250
2	D	35	85	0.175	0.425
3	F	30	115	0.150	0.575
4	B	24	139	0.120	0.695
5	H	20	159	0.100	0.795
6	G	15	174	0.075	0.870
7	I	12	186	0.060	0.930
8	C	8	194	0.040	0.970
9	J	5	199	0.025	0.995
10	E	1	200	0.005	1

Figure 1: The cumulative utility curve.

proportion of the utility can be gained by a substantially smaller data set proportion. Further, as the record proportion approaches 1 (that is, 100 per cent), the marginal (added) value decreases, possibly to a point where the increase is practically negligible.

The question of inequality may have important implications from the standpoint of data consumption – the use of customer data to support marketing decision-making. Should marketing efforts target the entire population, or rather focus on a subset? Accordingly, should a marketing analyst observe the entire data set (which, in many real-world contexts, could be extremely large), or focus on specific subsets of records? With customer data sets in which the magnitude of inequality is high, the assumption is that analysis would tend to focus on a small subset of the most profitable customers, rather than on the entire list.

An example in the field of direct-mail marketing is the ‘Pareto Curve’ as used in the earlier mention of list segmentation.²⁰ It takes full advantage of the equivalent of the cumulative utility curve (Figure 1(b)) based on the inequality of data values. Statistical analysis is used to predict purchase intent for all consumers or prospects and ranking the list in descending order of this estimated probability). The goal is to decide to whom to mail (referred to in,²⁰ as the ‘how deep to dip’ decision.) Such a decision is often supported by a curve that is, essentially, equivalent to (and typically looks like)

Figure 1(b), where the vertical axis reflects the ‘cumulative per cent of the purchase intent associated with the top rank-ordered names’, and the horizontal axis reflects ‘cumulative per cent of rank-ordered names sent an offer’. Indeed, an 80/50 point (capturing 80 per cent of the purchases by mailing only to the top 50 per cent of the rank-ordered list) is a frequent benchmark for a very successful list segmentation.

Modeling and measuring utility inequality in large data sets

For a large data set (large N), we represent the distribution of utility among the records as a random variable u with a known probability density function (PDF) $f(u)$. From the PDF we can calculate the mean $\mu = E[u]$, the cumulative distribution function (CDF) $F(u)$, and the proportion point function (PPF, the inverse of CDF), $G(p)$. In this study we demonstrate the computations for the continuous *Pareto distribution* – used later to analyze utility inequality in a customer database. Similar computations can be applied to virtually any other statistical distribution (for example, Uniform, Exponential, Weibull and Discrete). The Pareto distribution (Figure 2) is commonly used in economic, demographic and ecological studies. It is characterized by two parameters: Z , the minimum value and w , the decline rate. The highest probability is assigned to the lowest possible value of $Z > 0$ (Z can be arbitrarily close to 0). The probability

Figure 2: Pareto distribution: (a) PDF; (b) CDF and (c) PPF.

Figure 3: Obtaining the cumulative utility curve.

declines as Z grows and the parameter $w \geq 1$ defines the rate of decline:

$$\begin{aligned} f(u) &= wu^{-(w+1)} / Z^{-w}, u \geq Z; \\ F(u) &= 1 - (u/Z)^{-w}, u \geq Z; \\ G(p) &= Z / (1-p)^{1/w}; \mu = wZ / (w-1) \end{aligned} \quad (1)$$

To assess the extent to which records vary in their utility, we define R , the proportion of highest utility records, as a $[0,1]$ ratio between the N^* records of highest utility (that is, the top N^* when rank ordered in descending order) and N , the total number of records (for example, $R=0.2$ for a data set with $N=1\,000\,000$ records and $N^*=200\,000$ records that offer the highest utility). The *cumulative utility curve* $L(R)$ is a $[0,1]$ proportion of the overall utility as a function of R . $L(R)$ can be calculated from the proportion point function $G(p)$. For a large N , the added utility for a small probability interval $[p, p+\Delta p]$ can be approximated by $N \cdot G(p) \cdot \Delta p$ (Figure 3(a)).

Taking $\Delta p \rightarrow 0$, integrating the PPF over $[1-R, 1]$ (Figure 3(b)), and dividing the result by the total utility (approximated by μN) we get the cumulative utility curve $L(R)$ (Figure 3(c)):

$$L(R) = N \int_{1-R}^1 G(p) dp / N\mu = \int_{1-R}^1 G(p) dp / \mu \quad (2)$$

where,

R – The $[0,1]$ proportion of highest utility records;

$L(R)$ – The cumulative utility curve of the utility variable u , within $[0,1]$;

N – The number of data set records;

u, μ – The utility variable and its mean;

$G(p)$ – The PPF of the utility variable u .

The curve $L(R)$ is defined for $[0,1]$, where $L(0)=0$ and $L(1)=1$, and does not depend on N or on the utility unit. The curve is calculated by ‘backwards integration’ over $G(p)$, which is monotonically increasing; hence, it is

monotonically increasing and concave within $[0,1]$. The first derivative of $L(R)$ is therefore positive and monotonically decreasing, and the second derivative is negative. The maximum point of the curve (that is, $L(1)=1$) corresponds to the maximum possible data set utility u^D and the curve reflects the maximum portion of overall utility that can be obtained by the partial data set – that is, when only a portion R of the data set is available, the utility of $u^D L(R)$ can be achieved at best.

Using the cumulative utility curve $L(R)$, we next develop the inequality index (φ), which measures the relative area between the curve and the 45° line (that is, $f(R)=R$). This area is highlighted in Figure 3(c), and can be calculated by:

$$\begin{aligned}\varphi &= \left(\int_0^1 L(R) dR - \int_0^1 R dR \right) / \\ &\quad \int_0^1 R dR = \left(\int_0^1 \left((1/\mu) \int_{1-R}^1 G(p) dp \right) dR - 0.5 \right) / 0.5 \\ &= 2 \int_0^1 \left((1/\mu) \int_{1-R}^1 G(p) dp \right) dR - 1 \\ &= (2/\mu) \int_0^1 p G(p) dp - 1\end{aligned}\quad (3)$$

The value of φ is within $[0,1]$, where a higher value indicates a greater inequality. The lower bound, $\varphi \rightarrow 0$, indicates perfect equality – data set records with identical and deterministic utilities and a curve that approaches $L(R)=R$. The upper bound, $\varphi \rightarrow 1$, indicates a high degree of inequality – a small portion of records with a relatively high utility, while the utility of most other records is substantially lower. The corresponding curve in this case approaches $L(R)=1$ (with, technically, a vertical line at $R=0$, rising to $L(R)=1$).

The curve and the index can be further evaluated for specific distributions and can be often expressed using a closed analytical form. For the Pareto distribution, the

evaluations are:

$$\begin{aligned}L(R) &= \frac{1}{\mu} \int_{1-R}^1 G(p) dp = \frac{w-1}{wZ} \cdot \\ &\quad \int_{1-R}^1 \frac{Z}{(1-p)^{1/w}} dp = R^{1-\frac{1}{w}} \\ \varphi &= \frac{2}{\mu} \int_0^1 p G(p) dp - 1 = \frac{1}{2w-1}\end{aligned}\quad (4)$$

The Pareto curve and the index do not depend on the minimum value Z , but only on the decline rate w . Inequality decreases with w , where $w=1$ indicates the highest possible inequality ($L(R)=1$, $\varphi=1$). Conversely, with $w \rightarrow \infty$, $L(R) \rightarrow R$ and $\varphi \rightarrow 0$. The utility now is approximately identical for all instances: $f(u) \approx 1$, for $u \approx Z$ and ~ 0 otherwise (that is, $u \approx Z$ with probability ~ 1).

As noted earlier, the cumulative utility $L(R)$ is equivalent to the Lorenz curve and the inequality index is similar in nature to the Gini index – long-standing statistical tools for modeling and measuring inequality in large populations.

UTILITY INEQUALITY IN REAL-WORLD ALUMNI DATA

To illustrate assessment of utility inequality, we use data samples from a real-world customer database, which is used to manage alumni relations in a large university. This data resource is critical for the organization, as gifts by alumni, parents and friends account for a majority of its revenue. It is used by different departments for managing donors, tracking gift history and managing pledge campaigns. Similar systems exist in many organizations (for example, universities, museums, performing art organizations and charity funds) that depend on existing and potential donors for revenue.

Data samples

This study evaluates large samples from two key alumni data sets:

(a) *Profiles (358 372 records)* captures profile data on donors. Beside a unique identifier

(Profile ID), each record in this data set contains a large set of descriptive attributes, which were indicated by key data users as the ones most commonly used for managing alumni relations and/or for classifying profiles. Some attributes (for example, *Graduation Year* and *School*, *Gender* and *Ethnicity*) are typically included when a record is added to the data set, and are unlikely to change later, while others (for example, *Marital Status*, *Home Address*) are included when the record is added, but are likely to change over time. Some attributes are not included when a profile record is added, but are updated later (for example, *Income* and *Occupation*).

Another attribute, a 0/1 *Prospect* classification, splits the data set into two, reflecting two fundamentally different usages of the data and, accordingly, we analyze the two subsets separately. The organization labels 'prospects' (11 445 records, ~3 per cent of the data set) as donors who have made large contributions, or are assessed to have a potential for a substantial gift in the future. Prospects are not approached via regular campaigns. Instead, each prospect is assigned a staff member responsible for maintaining an ongoing contact with them (including, for example, invitations to special fund-raiser events and perks such as tickets to shows and sport events). Donors who are not classified as prospects (~97 per cent of the data set) are typically approached via pledge campaigns, each targeting a large donor base (for example, via phone, mail or E-mail).

(b) *Gifts* (1 415 432 records) captures the history of gift transactions. Beside a unique identifier (*Gift ID*), this data set includes a *Profile ID* (foreign key that links each gift transaction to a specific profile), *Gift Date*, *Gift Amount* and a few administrative attributes that describe payment procedures. Importantly, in this study we evaluate inequality in the *profiles* data set, while the *gifts* data set is used for assessing the *utility* of each profile.

The sample data in these data sets was collected between 1983 and 2006, and represents approximately 40 per cent of the data volume that is managed in the actual system. In 1983 and 1984, soon after system implementation, a bulk of records corresponding to earlier activities were added (203 359 profiles, 405 969 gifts), and since then, both data sets have grown gradually. The average annual growth of the *profiles* is 7044 records (sd: 475), while *gifts* grows by 45 884 records annually (sd: 6147). To guarantee confidentiality, some attribute values were masked in these samples (for example, actual addresses and phone numbers, graduation school, gender and ethnicity codes) and all gift amounts were multiplied by a constant factor.

Attributing utility and assessing inequality in profile records

The utility of using alumni data is reflected by the transactions in the gifts data set. Profile data along with past gift transactions, are used to identify and approach alumni with high donation potential. Gift transactions reflect the outcome of these efforts and can be linked to individual profile records. The assumption that future purchases (gifts, in this case), to a large extent, can be predicted by past activities is supported by the correlations between annual donation amounts and 'inclinations' (see Table 2). Inclination was coded as 1 for a profile if there was at least one donation (in *Gifts*) in the most recent 5 years (2002–2006) and 0 if not. The correlations between annual inclinations are positive and significant. The correlations between annual gifts are also positive and significant, but smaller. These scores are much lower for prospects compared to non-prospects.

Following these correlation results, we use the *average annual dollar amount gifted in the most recent 5 years* (2002–2006), identified from *Gifts*, as a proxy for the utility of profile records. Utility is 0 if a

Table 2: Correlations between annual inclinations and amounts*

		Prospects (11 445 records)					Non-prospects (346 927 records)				
		2002	2003	2004	2005	2006	2002	2003	2004	2005	2006
Inclination	2003	0.566	—	—	—	—	0.529	—	—	—	—
	2004	0.545	0.563	—	—	—	0.510	0.521	—	—	—
	2005	0.519	0.555	—	—	—	0.473	0.504	0.503	—	—
	2006	0.493	0.508	0.550	0.516	—	0.442	0.466	0.498	0.503	—
Amount	2003	0.240	—	—	—	—	0.399	—	—	—	—
	2004	0.199	0.139	—	—	—	0.359	0.389	—	—	—
	2005	0.157	0.061	0.146	—	—	0.301	0.351	0.412	—	—
	2006	0.016	0.010	0.062	0.020	—	0.271	0.315	0.341	0.386	—

*All correlations were obtained with P -Value ~ 0 .

person has made no donations in this time period, and positive otherwise. The utility distributions are shown in Figure 4(a). For non-prospects, the mean utility is US\$6.7, the standard deviation is \$38.1 and the proportion of profiles associated with 0 utility (that is, no gifts within the past 5 years) is very high (~ 88 per cent). For prospects, the mean and the standard deviation are substantially higher (\$1 303.5 and \$15 506, respectively) and the proportion of profiles associated with 0 utility (~ 54 per cent) is substantially lower.

The corresponding cumulative utility curves are shown in Figure 4(b). Assuming a Pareto distribution (equations 1 and 4), we used log-log regression to estimate the curves and the inequality indices (Gini indices). For non-prospects, the approximated curve is $L(R) = R^{0.111}$ (P -value: ~ 0 , Adjusted R^2 : 0.535). The equivalent Pareto parameter $w = 1/(1 - 0.111) = 1.124$, and the inequality index is $\phi = 1/(2w - 1) = 0.8$. The approximate curve for prospects is even steeper: $L(R) = R^{0.053}$ (P -value: ~ 0 , Adjusted R^2 : 0.546). The equivalent Pareto parameter $w = 1.056$, and the inequality (Gini) index is $\phi = 0.9$. In both cases, the Pareto distribution appears to be a reasonable fit for approximating the curves, although other asymmetric distributions (for example, Weibull or Exponential) may also fit the bill.

The inequality indices ($\phi = 0.8$ for non-prospects and 0.9 for prospects) suggest a high magnitude of inequality in gift-giving, both for prospects and for non-prospects. This has important business implications as it may suggest that a large portion of the data resource is underused – an opportunity for increasing gifts (indeed, 54 per cent of prospect records and 88 per cent of the non-prospect records are associated with 0 utility). It may also highlight the need to develop differential policies for managing records in this data resource. As further discussed in the following section, a better understanding of the business implications

Figure 4: Alumni profiles utility (a) Histogram and (b) Cumulative utility curve.

Figure 5: The (a) Utility; (b) Cost and (c) Net benefit curves.

of utility inequality requires recognition of data management costs and the possible utility/cost trade-offs.

INEQUALITY AND UTILITY/COST TRADE-OFFS

We have argued that inequality in the utility of data set records has important implications for data management. This can be evaluated from an economic perspective by assessing the effect of inequality on utility/cost trade offs and the overall net benefit. We consider u , the aggregated utility variable with corresponding maximum utility u^D and cumulative utility curve $L(R)$. We define $U(R)$, (the utility curve in Figure 5(a)), as the maximum possible utility as a function of R :

$$U(R) = u^D L(R) \quad (5)$$

where, $U(R)$ – The maximal possible utility as a function of R (the proportion of highest utility records); u^D – The maximal possible utility for the entire data set (that is, for $R = 1$); $L(R)$ – The cumulative

utility of the aggregated utility variable u , as a function of R .

The cumulative utility curve $L(R)$, hence $U(R)$, are monotonically increasing with a declining marginal return – a property critical for supporting our argument of utility/cost trade-offs. This property is explained by our definition of the proportion R as reflecting the sorting of records in descending utility order.

Managing data sets involves costs. We assume identical variable cost per record, uncorrelated to the record's utility. Accordingly, we initially model the variable cost as a linear curve (Figure 5(b)). This has a variable component c^v that is linearly proportional to the data set size (and, hence, to R), and a fixed component c^f that is independent of the data set size:

$$C(R) = c^f + c^v R \quad (6)$$

where, $C(R)$ – The data set cost for R (the proportion of highest utility records);

c^f c^v – Fixed cost and unit variable cost, respectively.

Assuming that utility and cost are scaled to the same monetary units, the net benefit contribution $B(R)$ of the data set is defined as the difference between utility and cost (Figure 5(c)):

$$B(R) = U(R) - C(R) = u^D L(R) - (c^f + c^v R) \quad (7)$$

Due to c^f $B(R)$ is negative at $R=0$ (the entire curve may be negative if $C > U$ for all R). It is concave and has a single maximum within $[0,1]$. An optimum, R^{OPT} can be obtained by comparing the first derivative of (8) to 0:

$$\begin{aligned} dB(R)/dR &= u^D (dL(R)/dR) - c^v = 0, \\ \text{or } dL(R)/dR &= c^v / u^D \end{aligned} \quad (8)$$

Below R^{OPT} the net benefit can be improved by increasing R , as the added utility is higher than the added cost. Beyond R^{OPT} the marginal cost exceeds the marginal utility and increasing R reduces the net benefit. For a steep curve (that is, $L(R) \rightarrow 1$, $\varphi \rightarrow 1$), or when the variable cost is significantly higher than the maximal utility (that is, $c^v \gg u^D$), the optimum approaches a low-record proportion (that is, $R^{OPT} \rightarrow 0$). If no positive R^{OPT} exists, the data set cannot provide a positive net benefit due to the fixed cost c^f . Conversely, if the variable cost is relatively low (that is, $c^v \ll u^D$), R^{OPT} is obtained at a high-record proportion (that is, $R^{OPT} \rightarrow 1$). With high equality (that is, $L(R) \rightarrow R$, $\varphi \rightarrow 0$), the solution will be at one of the edges – either near $R^{OPT}=0$, or near $R^{OPT}=1$. Notably, regardless of whether the R^{OPT} solution is within the $[0,1]$ range or at the edges, a positive net benefit is not guaranteed and has to be verified.

The optimality equation (8) can be extended for the Pareto distribution (equation 3):

$$\begin{aligned} dL(R)/dR &= (1 - 1/w) R^{-1/w} - c^v / u^D, \\ \text{and } R^{OPT} &= [(1 - 1/w) u^D / c^v]^w \end{aligned} \quad (9)$$

For $w > 1$, the optimum R^{OPT} for the Pareto distribution is always positive. It is within $(0,1]$ when the utility/cost ratio is $c^v / u^D \geq 1 - (1/w)$, otherwise, the optimal net benefit is obtained at $R^{OPT}=1$. The optimum approaches 0 for a high degree of inequality ($w \rightarrow 1$), that is, when the great majority of the utility is obtained from a relatively small number of records. The dependence of R^{OPT} on the utility/cost ratio grows with equality (that is, greater w). When the variable cost is very small (that is, $c^v \ll u^D$), the optimal net benefit is more likely to be obtained when the entire data set is included ($R^{OPT}=1$). When the variable cost is substantially large, the optimal net benefit is more likely to be at $R^{OPT} < 1$. For a high degree of inequality (that is, $w \rightarrow \infty$), the expression $(1 - 1/w)^w$ converges to a constant $1/e$. If the utility is higher than the variable cost ($u^D > c^v$), $(u^D / c^v)^w \rightarrow \infty$, and the optimum is obtained for the entire data set (that is, $R^{OPT}=1$). If $u^D < c^v$, $(u^D / c^v)^w \rightarrow 0$, $R^{OPT} \rightarrow 0$, and the data set is unlikely to yield any positive net benefit.

Utility/cost trade-offs for a Pareto distribution can also be assessed at the record level. Record utilities in this distribution are non-negative (equation 1), where Z represents the lowest possible utility (which can be arbitrarily close to 0). If the variable cost per record is lower than Z , the utility will always supersede the variable cost; hence, R will be maximized at 1 (it is still possible that the entire data set will not be implemented if the fixed cost is too high). On the other hand if the variable cost per record is greater than Z , the optimal R^{OPT} is likely to be lower than 1.

Importantly, real-world systems have capacity limits on the data volumes they can effectively process and store. Exceeding this capacity requires upgrading to a more powerful configuration – likely at a higher fixed cost and possibly at a different variable cost. This may require some adjustments

to the cost model – possibly representing it as a piece-wise linear curve. However, it can be assumed that even with such an adjustment, the cost curve will always be monotonically increasing (or non-decreasing) with volume. Therefore, the argument about the existence of a maximum net benefit point, which can possibly be internal to the evaluated range, still holds.

Implications for data usage and management

As discussed earlier, understanding the inequality in the utility of data may have important implications for data consumption. We further argue that assessing inequality and the associated utility-cost trade-offs may also inform different aspects of data administration:

(a) Data Acquisition and Retention Policies: The magnitude of utility inequality can impact the implementation of data acquisition and retention policies in customer data sets. When the inequality is low (that is, $L(R)$ converging to the 45° line and $\phi \rightarrow 0$), all records offer similar utility and the decision would be either to implement the same policy to the entire data set or not implement it at all. When inequality is high (that is, $L(R) \rightarrow 1$, and $\phi \rightarrow 1$), depending on utility/cost trade-offs, the designer may apply certain policies to high utility (or low-utility) records only, and manage them differently. A typical example of such differentiation in real-world data management is the retention of older data. Outdated customer records are often excluded from actively used data sets (and, perhaps, archived) and a utility cost assessment can help determining the optimal date range to be retained in the active data set.¹⁰ A similar argument holds for data acquisition policies – some agencies offer list enhancement at a certain cost per record. Inequality analysis can help deciding whether the entire customer data set should be enriched, or maybe just a subset of it.

(b) The Design of Data Environments: high differentials in the utility of data may affect the design of data environments, such as a data warehouse (DW) that manages large data sets.⁹ Managing large data sets requires higher investments in DW infrastructure (for example, more powerful database servers) and data delivery tools (for example, superior business intelligence and data analysis tools), as certain system configurations may limit the volumes of data that can be effectively managed. Investing in a powerful DW infrastructure will be harder to justify if a majority of the utility comes from a small fraction of the data that can be managed by a smaller/less expensive system.

(c) Data Quality Management: Differentiating utility can help define superior measurements for data quality dimensions (for example, completeness and accuracy), that reflect quality assessment in context.¹⁹ Further, differentiating the data records based on utility contribution can help prioritize and make quality management efforts more efficient. We next describe an evaluation of utility-cost trade-offs in the alumni data toward setting optimal data quality improvement priorities.

UTILITY/COST TRADE-OFFS IN MANAGING THE ALUMNI DATA

A possible application of utility-cost analysis is assessing data quality improvement alternatives, and identifying the best solution. Several studies (for example,^{21,22}) have underscored the importance of managing customer data at a high-quality level. Data quality defects (for example, missing, inaccurate and/or outdated data values) might prevent managers and analysts from having the right picture of customers and their purchase preferences and hence might damage marketing efforts significantly. Some studies (for example,^{2,23}) have also discussed methodologies and techniques for improving the quality of customer data.

Figure 6: Alumni profiles utility (a) Histogram and (b) Cumulative utility curve.

Here, we demonstrate how utility-cost analysis could be used for evaluating alternatives for improving the quality of the alumni data.²⁴ A preliminary evaluation of the profiles data set indicated major data quality issues. Approximately 84 per cent of the prospect profiles and 94 per cent of the non-prospect profiles are missing data values in key attributes, including some that are crucial for alumni-relation management and solicitation efforts (for example, Income, Profession, Home and Business Address). Further, in approximately 22 per cent of the prospect profiles and 50 per cent of the non-prospect profiles, that data had not been audited and updated within the past 5 years. This implies that a large proportion of the profiles data set has outdated and possibly inaccurate data (for example, due to changes of address, marital status, income level and other important alumni attributes).

The vast majority (~97 per cent) of alumni profiles is classified as non-prospects, and these are associated with relatively low contribution (~88 per cent of the non-prospect alumni have made no contribution within the past 5 years). The alumni managers link the large proportion of zero-utility profiles to the high rate of data quality defects. In the evaluation that we demonstrate here, we assume that the data subset targeted for quality improvement includes profiles of alumni who have graduated within the past 30 years, which have some data quality defects (that is, not updated within the past 5 years and/or with

missing values) and have no associated utility within the past 5 years. The total number of records in this data set is 174 356, and the number of records declines with *Record Age* (Figure 6(a)) – a variable that defines the number of year from when the record was added to the database (the year of graduation) to the point of evaluation (for example, the age of a record that was added last year would be 1).

To assess the utility contribution potential of the targeted profiles – we evaluated the utility (average annual contribution within the past 5 years) associated with non-prospect alumni who have made some contribution, per record age (Figure 6(a)). While the number of record decreases with record age, the utility increases – alumni who have graduated many years ago would typically have higher income and financial resources, and would be willing to make higher contributions than recent graduates.

Assuming a Pareto distribution, the estimated cumulative utility curve (Figure 6(b)) for the targeted subset of profiles is $L(R) = R^{0.858}$, and the corresponding inequality index is $\phi = 0.166$. The calculation followed these steps:

- The data set proportion variable R corresponds to the number of years (out of 30) and the number of associated profiles (out of 174 356 records). As the less-recent profiles have higher contribution potential, it would be reasonable to improve the data quality of profiles with high record

age first, and go 'backwards' to the more recent profiles. For example – the R corresponding to record age 30 (4831 profile records) is $4831/174\,356 = 0.028$, the R corresponding to profiles of age 29 and 30 years ($4831 + 5334$ records) is 0.058, and so on.

- To estimate the utility-contribution potential per record age, we multiplied the estimated utility by the number of targeted records – for example, the potential utility of age 30 years is $4831 \times 74.81 = 361\,415$, and the overall annual utility potential was estimated at 8362030. Using this estimation, $L(R)$ was calculated as the cumulative utility proportion for each R – for example, $L(0.028) = 0.043$, $L(0.058) = 0.095$ and so on.
- We used log-log regression (F -Value = 2369, P -Value = ~ 0 , Adjusted $R^2 = 0.987$) to estimate the curve, based on the 30 points (one per record age).

We now use this curve to evaluate two possible data quality improvement treatments: (a) *Alumni Survey*: the survey will be mailed to the targeted alumni and the surveyed person will be asked to update his personal details. The response rate of such a survey is estimated at 30 per cent, and the average cost per record is estimated at \$8 (a maximum variable cost of $C^v = \$1\,394\,848$), including printing and mailing cost and the time needed to handle the delivery and update the database. The survey also involves some fixed costs (for example, campaign planning and initiation, managerial overhead), but they are relatively small and considered negligible for the matter of our analysis. (b) *Comprehensive Investigation*: updating data on a person can be done by searching the web, hiring external agencies or assigning a contact person. Such methods have been commonly applied to prospect profiles, but not to non-prospects due to the high cost. A typical cost of such an investigation would be \$26 per record (a maximum variable

cost of $C^v = \$4\,533\,256$, plus some negligible fixed costs), and the success rate is estimated at 90 per cent.

The evaluation assumes that 20 per cent of the alumni with corrected data will make donations within the next 3 years at an annual rate similar to the average annual contribution of non-prospect alumni with the same record age, who made some donations. Accordingly, the maximum contribution potential for the alumni survey is estimated at $U^D = \$943\,420$ and the optimum R is at $R^{OPT} = 0.580$ (equation 9) – equivalent to the subset of profiles with record age between 13 and 30 years. The corresponding maximum net benefit (equation 7) is $B = \$134\,408$. For the comprehensive investigation, the maximum utility-contribution potential is estimated at $U^D = \$1\,749\,525$. The optimum is at $R^{OPT} = 0.331$ (equation 9) – equivalent to the subset of profiles with record age between 20 and 30 years. The corresponding maximum net benefit (equation 7) is $B = \$249\,017$.

As the maximum net benefit of the second treatment is higher, the recommendation would be to run a comprehensive investigation for all profiles with records age between 20 and 30 years. However, some additional net benefit can be gained by surveying alumni with profiles of record age between 13 and 19 years (corresponding, approximately, to $0.331 \leq R \leq 0.580$), as within this range the marginal utility per record is higher than the variable cost. The estimated added utility within this age is $943\,420 \times (0.580^{0.858} - 0.331^{0.858}) = \$360\,245$, the estimated added cost is $1\,394\,848 \times (0.580 - 0.331) = \$347\,317$, and the added net benefit is \$12928.

Our analysis here highlights the possible need for differentiating policies when it comes to improving the quality of customer data. Some treatments are relatively expensive, hence, should be applied only to a small subset of the customer profiles – those with higher contribution potential that can justify

the cost. For example, according to this analysis, it would be recommended not to apply any of the analyzed treatments for profiles of record age between 1 and 12 years. Obviously a real-world application will require a more thorough evaluation and more precise estimations of utility and costs (for example, by soliciting the numbers from knowledgeable managers, or surveying vendors who specialize in customer list enhancements). The alumni managers indicated other possible data quality improvement treatments (for example, E-mail surveys, automated search in public databases) that can be analyzed in the same manner and can possibly fit the entire record age range.

CONCLUSIONS

Our study offers a novel perspective toward a more cost-effective usage and administration of customer data. While encouraging the need to include economic considerations into data management, the study demonstrates the insights gained by modeling and analyzing inequality in the utility of customer data, using an adaption of the Lorenz curve and the Gini index – commonly used statistical tools for understanding inequality in large populations. Modeling and measuring utility and its inequality can serve different purposes, as in the process of analyzing inequality, economic trade-offs are linked to the distribution of utility in large data sets. By assessing utility and its distribution, data administrators can now consider a refined and differentiating treatment of data records within a data set instead of treating the data set as a single entity. Assessing the magnitude of inequality can help determine the current state of a data resource, identify improvement targets and track the progress of improvement efforts directed through periodic evaluation. Further, it may help detecting and preventing over-investments in data with low utility. Inequality assessment may also direct better usage of

data, as it highlights data subsets that are associated with higher (or lower) utility. It can hence indicate incorrect use of certain data sets, and guide exploration and experimentation of alternative usages.

The study is not without limitations. To get a true picture of economic trade-offs, utility and cost must be assessed for the entire data resource, not just for a subset, as we have demonstrated here. Generally, the costs considered and discussed in this study are associated directly with data management. However, managing customer relations in real-world settings involves other costs, such as, costs associated with maintaining contact with customers (for example, mailing, phone calls) and retaining them (for example, promotions, loyalty cards). Though not associated directly with managing the data, these costs may significantly affect customer data-management decisions and should be explored.

We develop inequality measurements assuming a Pareto distribution. Although this choice appears reasonable for modeling donation behavior (as shown in our example), other statistical distributions may better reflect other real-world data usage scenarios. Our cost model assumes an equal variable cost per data set record. Variable costs may not always be linear with number of records (for example, purchasing bulks of customer data for list enhancement at a discounted price).

The current model considers a static ‘snapshot’ of utility. Utility contribution may dynamically change over time and its effect on data management decisions must be modeled differently. Importantly, inequality assessment alone does not provide a full picture of the current state of data resources. For instance, measurements of data quality which reflect the presence of defects are also important. However, assessing the inequality in data utility may offer insights that can prioritize data management efforts and ensure the ‘biggest bang for the buck’.

REFERENCES AND NOTES

- 1 Davenport, T.H. (2006) Competing on analytics. *Harvard Business Review* 84(11): 99–107.
- 2 Roberts, M.L. and Berger, P.D. (1999) *Direct Marketing Management*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.
- 3 Petrison, L.A., Blattberg, R.C. and Wang, P. (1997) Database marketing: Past, present, and future. *Journal of Direct Marketing* 11(4): 109–125.
- 4 Berger, P.D., Weinberg, B.D. and Hanna, R. (2003) Customer lifetime value determination and strategic implications for a cruise-ship company. *Journal of Database Marketing and Customer Strategy Management* 11(1): 40–52.
- 5 Berger, P.D., Eechambadi, M., Lehmann, G.D., Rizley, R. and Venkatesan, R. (2006) From customer lifetime value to shareholder value: Theory, empirical evidence, and issues for further research. *Journal of Services Research* 9(2): 156–167.
- 6 Blattberg, R.C., Malthouse, E.C. and Neslin, S.A. (2009) Customer lifetime value: Empirical generalizations and some conceptual questions. *Journal of Interactive Marketing* 23(2): 157–168.
- 7 Fader, P. and Hardie, B. (2009) Probability models for customer-base analysis. *Journal of Interactive Marketing* 23(1): 61–69.
- 8 Gupta, S. (2009) Customer-based valuation. *Journal of Interactive Marketing* 23(2): 169–178.
- 9 Even, A., Shankaranarayanan, G. and Berger, P.D. (2006) The data-warehouse as a dynamic capability: Utility-cost foundations and their implications for economics-driven design. in Proceedings of the International Conference on Information Systems (ICIS), Milwaukee, WI, USA.
- 10 Even, A., Shankaranarayanan, G. and Berger, P.D. (2007) Economics-driven data management: An application to the design of tabular data sets. *IEEE Transactions on Knowledge and Data Engineering* 19(6): 818–831.
- 11 Easterfield, T.E. (1964) Optimum variety. *Operations Research Quarterly* 15(2): 71–85.
- 12 Schechtman, E., Yitzhaki, S. and Artsev, Y. (2008) Who does not respond in the household expenditure survey: An exercise in extended Gini regressions. *Journal of Business and Economics Statistics* 26(3): 329–344.
- 13 Raileanu, K.E. and Stoffel, K. (2004) Theoretical comparison between the Gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence* 41(1): 77–93.
- 14 Beteille, A. (ed.) (1983) *Equality and Inequality: Theory and Practice*. New Delhi: Oxford University Press.
- 15 Anand, S. and Sen, A. (1995) Gender Inequality in Human Development: Theories and Measurement. Background Paper for the Human Development Report, United Nations Development Programme.
- 16 Lorenz, M.O. (1905) Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 9(9): 209–219.
- 17 Gini, C. (1912) Variabilità e mutabilità. Reprinted in E. Pizetti and T. Salvemini (eds.) *Memorie di metodologica statistica*. Rome: Libreria Eredi Virgilio Veschi.
- 18 Ahituv, N. (1980) A systematic approach towards assessing the value of an information system. *MIS Quarterly* 4(4): 61–75.
- 19 Even, A. and Shankaranarayanan, G. (2008) Comparative analysis of data quality and utility inequality assessments. In Proceedings of the European Conference on Information Systems (ECIS), Galway, Ireland.
- 20 Magliozzi, T. and Berger, P.D. (1993) List segmentation strategies in direct marketing. *OMEGA: International Journal of Management Science* 21(1): 61–72.
- 21 Coutheux, R.J. (2003) Marketing data analysis and data quality management. *Journal of Targeting, Measurement and Analysis for Marketing* 11(4): 299–313.
- 22 Foss, B., Henderson, I. and Johnson, P. (2002) Managing the quality and completeness of customer data. *Journal of Database Marketing* 10(2): 139–158.
- 23 Heinrich, B., Kaiser, M. and Klier, M. (2009) A procedure to develop metrics for currency and its application in CRM. *ACM Journal of Data and Information Quality* 1(1), 5.1–5.28.
- 24 The evaluation demonstrated in this section was done as a post hoc analysis, after the research project was completed. The cost numbers that we state have not been evaluated precisely, but are based on estimates by the alumni database managers.