Forthcoming, *Information Systems Research*

Reidentification Risk in Panel Data: Protecting for *k*-Anonymity

Shaobo Li, Matthew J. Schneider, Yan Yu, Sachin Gupta[1]

September 1, 2022

[1] Shaobo Li is Assistant Professor (shaobo.li@ku.edu), School of Business, University of Kansas, 1654 Naismith Drive, Lawrence, KS 66045; Matthew J. Schneider is Associate Professor (mjs624@drexel.edu), LeBow College of Business, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104; Yan Yu is Joseph S. Stern Professor of Business Analytics (yan.yu@uc.edu), Carl H. Lindner College of Business, University of Cincinnati, 2906 Woodside Drive, Cincinnati, OH 45221; Sachin Gupta is Henrietta Johnson Louis Professor of Management, (sg248@cornell.edu), SC Johnson College of Business, Cornell University, Ithaca, NY 14853.

Reidentification Risk in Panel Data: Protecting for *k*-Anonymity

Abstract

We consider the risk of re-identification of panelists in marketing research data that are widely used to obtain insights into buyer behavior and to develop marketing strategy. We find that 17% to 94% of the panelists in 15 frequently bought consumer goods categories are subject to high risk of reidentification through a potential record linkage attack based on their unique purchasing histories, even when their identities have been anonymized. We first demonstrate that the risk of reidentification is vastly understated by unicity, the conventional measure. Instead, we propose a new measure of reidentification risk, termed sno-unicity, that accounts for the longitudinal nature of panel data and show that it is much larger than unicity. To protect the privacy of panelists we consider the well-known privacy notion of *k*-anonymity, and develop a new approach called *graph-based minimum movement k-anonymization* (*k-MM)* that is designed especially for panel data. The proposed *k*-MM approach can be formulated as an optimization problem where the objective is to minimally distort variables in the original data based on weights which users pre-specify corresponding to their use case. We further show how our approach can be extended to achieve *l*-diversity. We apply the *k*-MM approach to two different panel datasets that are widely used in marketing research. To achieve a given privacy level, compared to several benchmark protection methods, the protected data from our method result in the least distortion in inferences about key marketing metrics such as brand market shares, share of category requirements, brand switching rates, and marketing-mix parameters estimated from a hierarchical Bayesian brand choice model.

Keywords: brand choice, data privacy, data sharing, hierarchical Bayesian model, optimization, unicity

1 Introduction

Individual-level panel data are commonly used in many areas of marketing research. For instance, web browsing data from Comscore's panel of over 2 million work and home panelists are widely used for measuring digital audiences (Lipsman et al., 2012). Similarly, Nielsen's TV and Radio panels are used for measuring viewership and listenership, respectively. In the consumer-packaged goods (CPG) industry, IRI and Nielsen's household panels provide purchasing data that are analyzed by retailers and manufacturers to develop marketing programs that drive brand and category performance. These companies are required to provide panelists with privacy agreements that explain how the data will be used and how the company complies with data privacy laws and regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Under these regulations, companies are required to anonymize personal identifying information (PII) such as first and last name, home postal address, or email address before any type of data analysis.

Although privacy laws are intended to reduce the leakage of personal data, individuals may still be subject to reidentification through data linkage, which is also known as a background attack. One reason is that the availability and type of external information used in such an attack can never be assessed with certainty (Finck and Pallas 2020), and non-PII attributes can be used for linking and reidentification. Sweeney (2000) showed that 87% of individuals in the United States have a unique combination of three attributes: gender, date of birth and 5-digit zip code, hence they can be re-identified. Other examples show that demographic information is not even necessary to destroy the anonymity of a data set. The New York State Taxi and Limousine Commission released the "anonymous" details of 173 million driving routes without PII or demographic information, but it was discovered that the driving patterns revealed the drivers' home addresses (Hern 2014). Narayanan and Shmatikov (2008) successfully de-anonymized users in an anonymized dataset of movie ratings of 0.5M subscribers of Netflix, released publicly by Netflix as part of a competition. In this case the background or external information of individuals was obtained from an IMDb dataset (www.imdb.com), which was also public. De Montjoye et al. (2015) illustrated that only four spatiotemporal points are sufficient to uniquely re-identify 90% of anonymized individuals in a credit card dataset.

Quasi-identifiers (QIDs, Sweeney 2000, 2002a) are attributes that are common between the dataset that is subject to privacy risk (the target dataset) and the external data and can be used to link them. QIDs do not directly identify individuals but may do so indirectly. In this paper, we follow this convention and assume that an intruder attempts to reidentify a target through record linkage by matching QIDs (Fung et al. 2010; Li and Sarkar 2011). The proportion of individuals with unique QIDs in the target dataset, which following De Montjoye et al. (2013) we call "unicity", is a natural and conventional measure of reidentification risk in the data privacy literature (Lambert 1993; Sweeney 2002; Li and Sarkar 2006; El Emam and Dankar 2008). In other words, an individual is said to be re-identifiable if she has a QID that is different from anyone else in the target dataset, regardless of whether an intruder's external data contains this individual's QID. As a measure of reidentification risk, unicity does not and should not depend on any particular intruder's knowledge because such information is hard to assess and is not known by data providers a priori. Therefore, unicity can also be considered a property of the target dataset for given QID attributes.

To reduce reidentification risk as measured by unicity, the seminal work of Sweeney (Sweeney 2002a) introduced k-anonymity, a privacy protection model which has been widely adopted in the literature and in practice, e.g., Google Cloud (https://cloud.google.com/dlp/docs/compute-k-anonymity). A released dataset is said to be k-anonymous if any individual in the dataset is indistinguishable from at least k-1 other individuals with respect to the QIDs. Therefore, the parameter k defines the degree of privacy and can be interpreted as follows: the probability of an individual being re-identified in the target data is at most $1/k$.

Most existing research on k-anonymity concerns reidentification in cross-sectional microdata, where each row typically consists of an individual's demographic information, which forms the QIDs, and information such as disease diagnosis or prescription, which is the sensitive information to be learned upon reidentification. Different from this literature, we are concerned with the reidentification risk in panel data, where each individual has multiple records across time. To our knowledge, reidentification risk and k-anonymization of panel data has not been well studied in the literature. We find that the reidentification risk in panel data is potentially much higher than indicated by conventional measures that were developed for

cross-sectional data. Further, we propose a new method to achieve *k*-anonymity in panel data since existing approaches for cross-sectional data cannot be directly applied to panel data.

In this paper we show two empirical applications to panel data that are widely used in marketing research and find that in both the risk of reidentification of panelists is extremely high. Our first application is to household panel data that report consumer packaged goods (CPG) purchases. In the US market the two leading providers of these data – IRI and Nielsen – jointly operate a National Consumer Panel (https://www.ncppanel.com/what-we-do/) which consists of about 120,000 households. Our second application is to data from a panel of physicians whose prescription-writing behaviors are tracked over time to evaluate the effectiveness of marketing activities such as detailing by sales representatives (for typical applications of such data see Manchanda et al. (2004), Liu et al. (2016), and Kappe and Stremersch (2016)). In this paper, we primarily focus on the household panel data and use the physician panel data to demonstrate the generality of our method.

The data we use in our household panel application were provided by IRI, which is responsible for data protection and privacy under both GDPR and CCPA (https://www.iriworldwide.com/en-us/company/global-privacy-statement/privacy-policy). Because of their commercial value, the anonymized household panel data gathered by market research companies are shared extensively with users like manufacturers and retailers, as well as third parties who offer data analytics and consulting services to the users (Bucklin and Gupta 1999). The purchasing data may also be combined with other information about the panelists, such as online browsing behaviors, media consumption, advertising exposures, participation in loyalty programs, and so forth. IRI protects the identities of participating households by exercising common precautions like removal of personal identifying information (PII) such as first and last name, home postal address, or email address. However, other personal information such as purchasing activities can be used to identify household members in the event of a data breach or a linkage attack. The question we address in this paper is the extent to which the identities of panelist households are at risk of disclosure, and how this reidentification risk can be reduced while minimally changing the data.

1.1 A Motivating Example

Figure 1 illustrates how individuals in the household panel data can be reidentified and the resulting loss of privacy. Imagine an "intruder" who has access to the household panel data and certain external data (labeled "Intruder's knowledge" in Figure 1). The goal of the intruder is to reidentify one or more anonymous panelists from the panel data, who are also in the external data, and both data share some common attributes (QID). Because panel data are gathered in a panelist-centric manner, they usually contain rich information about panelist households such as purchases in multiple product categories and at multiple retailers, while the intruder's knowledge is limited (e.g., purchases in a single category, such as salty snacks, from one retail store) but may contain individuals' true identities. The purchase records of salty snacks can then serve as the QIDs that the intruder can match with purchases of salty snacks in the panel data. An individual in the panel data is re-identifiable if at least one of her purchase records of salty snacks (QID) is unique, so that the match to external data can lead to reidentification, i.e., the panelist ID (PANID) in the household panel data is linked to the true identity in the external data. This is known as identity disclosure risk (Duncan and Lambert 1989), which is also the focus of this paper.

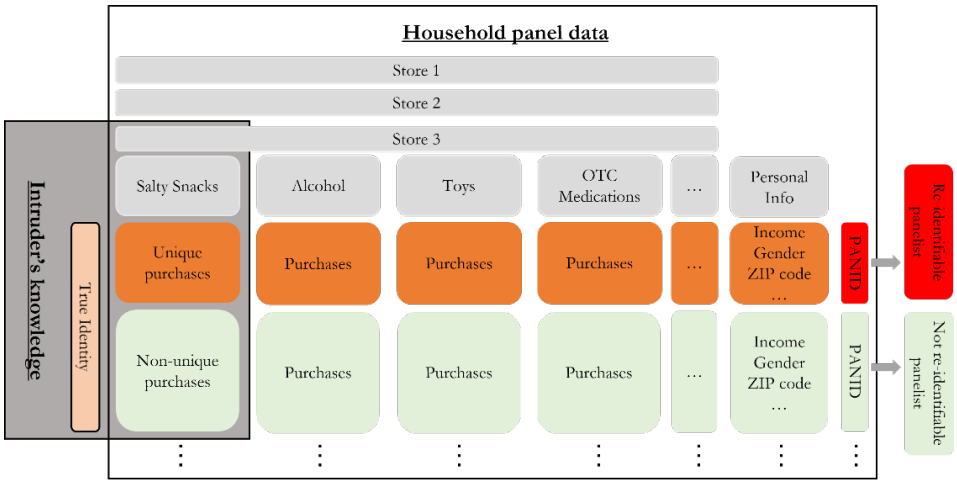

Figure 1. Illustration of the reidentification risk for household panel data

The resulting privacy loss is very consequential, since for the reidentified individual the intruder uncovers purchases in potentially sensitive categories such as alcohol, tobacco, pregnancy tests, or over-the-counter drugs (cross-category attack), purchases from competitor retailers (cross-retailer attack), and personal information such as household income, home address, zip code, and number of family members (personal

information attack), that is typically available in the household panel data. In fact, we found in our empirical analysis that among households that were re-identifiable based on their purchases of salty snacks in a given store, 20% bought beer and 2% bought cigarettes from a different store. These consequential privacy risks are known as attribute disclosure risk (Duncan and Lambert 1989). Finally, we stress that even if no sensitive information is revealed about the reidentified individual, the very act of reidentification is still a compromise of data privacy (Lambert 1993) and violates privacy laws and regulations.

Examples of intruders can be employees of retail chains or consulting companies, or individuals who have access to multiple data sources that are sufficient to compile a sample of transaction data from one retail store (intruder's knowledge) that contains individuals who are also present in the household panel data. For instance, an employee of a retail chain has access to some transactions data from the retailer's point-of-sale system, which may also include buyers' identities (e.g., credit card purchases or transactions linked to loyalty program databases). The "intruder" can also be a panelist's neighbor, colleague, friend or family member who does not intentionally wish to reidentify this panelist but knows one or more purchases made by this panelist which happen to be unique in the panel data. It is important to note that the intruder's knowledge is never known to the data provider a priori and is hard to estimate. A data vendor such as IRI or AC Nielsen must assess the reidentification risk based on the characteristics of the released data, not intruders' knowledge.

Table 1. Example of shopping transaction-level household panel data.

Transaction	Panelist ID	WEEK	Units of Lays bought	Units of Ruffles bought	QID
A1	A	2	2	2	(2, 2, 2)
A2	A	2	0	1	(2, 0, 1)
A3	A	2	2	0	(2, 2, 0)
B1	B	2	2	0	(2, 2, 0)
B2	B	2	2	1	(2, 2, 1)
B3	B	3	0	2	(3, 0, 2)
C1	C	2	0	1	(2, 0, 1)
C2	C	2	2	1	(2, 2, 1)
C3	C	3	1	0	(3, 1, 0)
D1	D	3	1	0	(3, 1, 0)
D2	D	3	0	2	(3, 0, 2)

We consider uniqueness-based reidentification risk or unicity, which in this case is defined as the proportion of buying households in the category who have at least one unique QID. However, we argue that unicity can underestimate the uniqueness of QIDs in panel data. To see this, consider Table 1 which shows

hypothetical transaction-level purchases of potato chips by four panelists identified pseudonymously (i.e., their names have been replaced with reversible pseudo-IDs) as A, B, C, and D. In this simplified example, we assume there are only four panelists and two brands (Lays and Ruffles). The column labeled "QID" is a quasi-identifier, a vector constructed as the combination of the week and number of units purchased of the two brands. To determine individual re-identifiability, the data provider first filters all households in the panel data that have unique QIDs. For instance, in Table 1 the QID (2, 2, 2) is unique among the 11 observations, which implies that household A is re-identifiable.[2] Except for transaction A1, all the other transactions are not unique, hence the reidentification risk is computed using unicity is one out of four panelists, or 25%.

However, this calculation disregards an important characteristic of panel data. If individual A is re-identified, then all transactions made by A should not be considered and need to be removed. After removal of transactions A1, A2 and A3 we find that transaction B1 is unique, hence individual B is re-identifiable. Similarly, individuals C and D are both subsequently found to be re-identifiable. Therefore, the reidentification risk is 100% and not 25%. We emphasize that we say an individual is "re-identifiable" rather than "re-identified" because the intruder's knowledge is not relevant in this computation. We call this measure of reidentification risk "sno-unicity" (sno for snowballing) because it is computed recursively. Sno-unicity can be thought of as a worst-case scenario reidentification risk, which is realized when the intruder knows all the unique QIDs as illustrated in this example.

It is clear that reidentification risk depends heavily on the uniqueness of the QIDs in the data. Unlike many existing studies where individuals' (time-invariant) demographic characteristics are QIDs, we focus on the purchase data as the QIDs, because the purchase records often vary within each individual, and the goal of our study is to anonymize such heterogeneous records of each individual. Clearly, inclusion of both demographics and purchases as QIDs will increase the risk of reidentification beyond the levels we report in this paper, and therefore make the case for privacy protection even stronger. Further, we focus on identity disclosure risk and do not assume specific sensitive attributes.

[2] In this paper we consider a simplified scenario in which the intruder matches one QID per household at a time rather than multiple transactions because the number of transactions in the external data often do not match that in the panel data.

1.2 Summary of Contributions

In this paper we make three main contributions. First, we demonstrate that based on unicity the reidentification risk in household panel data (physician panel data) is very high based on a QID that is composed of variables indicating time and units bought (prescriptions written) of the largest selling ten (four) brands.[3] In Figure 2 we show the unicity of the panel data in each of 15 frequently purchased consumer packaged goods. We find that the unicity ranges from 14.4% for mayonnaise to as high as 64% for carbonated beverages, implying that 64% of panelist households are re-identifiable based only on their carbonated beverage transactions. In the prescription data 57% of physicians in the panel are re-identifiable based on unicity.

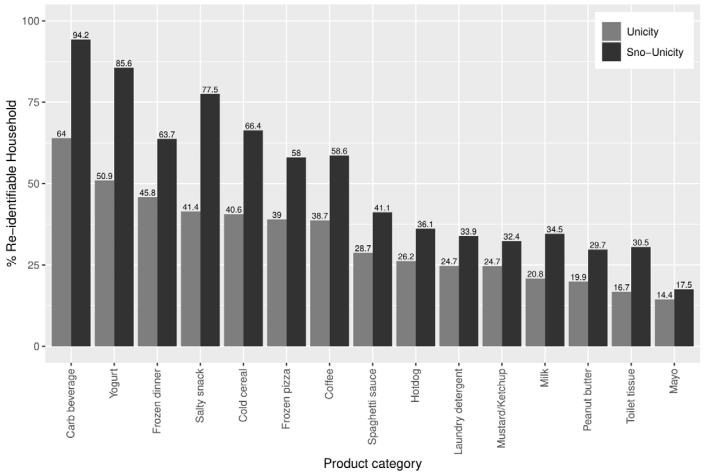

Figure 2: Reidentification risk in 15 product categories in IRI data. The quasi-identifier (QID) is composed of transaction week and number of units purchased of largest selling ten brands in the category.

Second, we argue that the conventional measure, unicity, understates reidentification risk because it does not account for the longitudinal nature of panel data, namely, that there are multiple records per panelist. To account for this characteristic of the data, we propose a new measure of reidentification risk called snowballing unicity (we abbreviate it as sno-unicity) which captures the potentially recursive nature of reidentification of panelists in longitudinal data. Based on sno-unicity we find that the reidentification risks in both household scanner panel data and the physician panel data are significantly higher than those based on unicity. Figure 2 shows that sno-unicity is higher than unicity in all 15 categories in the household panel data.

[3] Subsequently in the paper we provide more details about the household panel data and the physician panel data, and the measures.

Alarmingly, over 94% of carbonated beverage customers are re-identifiable. We note that if instead of a single category, purchases in multiple product categories are included in the QID, the reidentification risk can only be higher because the resulting QID is even more likely to be unique. In the physician panel data, we find that the sno-unicity is 100%, implying that all physicians are potentially re-identifiable.

We further derive an individual-specific probabilistic measure of reidentification risk based on two byproducts from the sno-unicity computation; these byproducts are the number of unique QIDs of an individual, and the iteration at which an individual becomes re-identifiable. Application of this measure to the household panel data generates the insight that heavier category buyers and less brand loyal buyers suffer higher reidentification risk.

Our third contribution is to propose a new method to reduce reidentification risk in panel data by guaranteeing *k*-anonymity while ensuring minimal distortion. As mentioned earlier, existing approaches for *k*-anonymization are not readily applicable to panel data. Our proposed approach is designed for protecting panel data and is called *graph-based minimum movement k-anonymization* (we abbreviate this as *k-MM*). The term graph-based refers to the fact that the problem and its solution can be graphically represented, as we show in Sections 3 and 4. Our proposed approach offers the following advantages over existing approaches.

First, our approach generates protected data that preserve the dimensionality and variable types in the original data. This is unlike generalization and suppression (Sweeney 2002b), which are commonly used jointly for *k*-anonymization. Generalization, also known as domain generalization hierarchy, entails replacing the values of the QID attributes with more general values that represent a group of individuals with respect to those attributes. For example, the specific age of an individual can be generalized into intervals such as < 20, 21-30, 31-50, and > 50 years; this may result in substantial loss of utility due to the change of attribute type (Gelman and Park 2009). Suppression of data entails completely removing some records that do not satisfy *k*-anonymity. As a result, suppression either reduces the size of the protected data or creates missing values, which also affects the data utility adversely. Our approach is based on alteration, which means that we only *alter* the values in some cells of the target data if necessary. In Table 1 row 1, for example, we could

change the units bought of Ruffles from 2 to 1, so that 2-anonymity is achieved. As a result, the number of records for each individual and the type of the altered variable remain the same as in the original data.

Second, our approach achieves minimum distortion by solving an optimization problem where the objective function is the cost of distortion represented by distance moved (e.g., the cost of changing 2 to 1 is lower than changing 2 to 0). The problem can be elegantly represented in a graph where each vertex is an individual and each edge is a unique QID value (e.g., "2-2-2" in Table 1 row 1). Such a graph representation makes our *k*-anonymization problem easy to understand and interpret. Our graph is different from those used for studying computational complexity of *k*-anonymity (Meyerson and Williams 2004; Aggarwal et al. 2005), wherein one record per individual is assumed. Following our graph representation, the optimization problem can be formulated as an integer linear program, and we propose a heuristic based on divide-and-conquer to solve it efficiently. Further, unlike generalization, our formulation does not require any pre-specified domain generalization hierarchy because we alter values instead of generalizing to a higher level.

Third, the goal of minimal distortion of the original data is to preserve the accuracy of analyses that are subsequently performed on the protected data, without specifying a priori what those analyses might be. However, if the user wishes to preserve specific kinds of information in the original data, our method also allows the user to specify different levels of distortion for different attributes contained in the QID by incorporating weights. For instance, as we elaborate in the subsequent empirical application, if the household panel data are used for pricing decisions, the user may wish to incur less distortion in the purchases of larger market share brands. In a healthcare setting where demographic characteristics of patients are QIDs, decisions about vaccination policy may require that distorting patient age is more costly than distorting the patient's geographic location. In the empirical application we demonstrate that our protection approach can successfully achieve such a goal.

In both empirical applications we find that compared with benchmark methods, our proposed approach preserves the utility of the unprotected data much better on key measures of interest to the marketing manager: market shares of brands, brand loyalty measured as share of category requirements, brand

switching, and parameter estimates from a hierarchical Bayesian brand choice model (e.g., Allenby and Rossi 1998; Bruno et al. 2018).

The rest of this paper is organized as follows: Section 2 reviews related work on privacy protection. Section 3 discusses in greater depth data privacy concerns in panel data and formally defines our proposed measure of reidentification risk. We present the proposed data protection method in Section 4 and empirical results in Section 5. We conclude in Section 6 and discuss limitations and future research opportunities.

2 Related Work

The study of data privacy originated in the statistical disclosure limitation literature (Cox 1980; Duncan and Lambert 1986; Rubin 1993; Reiter 2005), which examines statistical agencies' attempts to release microdata for public access while controlling the risk of individual identity disclosure. Popular methods employed by agencies such as the U.S. Census Bureau include aggregation, suppression, top/bottom-coding, rounding, swapping, noise addition, and synthetic data; see, e.g., Matthews and Harel (2011) for a thorough review. In recent years, issues of data privacy have been actively studied in business domains including marketing and information systems, using consumer behavior, organizational, ethical, and economic perspectives (Smith et al. 1996; Malhotra et al. 2004; Goldfarb et al. 2012; Tucker 2014). Comprehensive reviews can be found in Smith et al. (2011), Martin and Murphy (2017) and Wieringa et al. (2021). Ferrell (2017) notes the gap between understanding privacy and taking actions to determine risk and protect privacy in marketing practice. Wedel and Kannan (2016, page 114) emphasize that research "needs to focus on how customers' privacy can be protected in the use of rich marketing data while maximizing the utility that can be derived from it by developing models and algorithms that can preserve or ensure consumer privacy."

Among the early papers that develop privacy protection methods for marketing research data, Schneider et al. (2017) developed a Dirichlet-Multinomial model based on ϵ-differential privacy (Dwork 2006, Machanavajjhala et al. 2008) for a company to protect its customer-level segment membership information when entering a data-sharing arrangement with another organization. Schneider et al. (2018) studied point-of-sale data and finds that sales data are very informative for uncovering retail store identities. They proposed a Bayesian approach to generate synthetic sales data that significantly reduce the reidentification risk while

preserving useful information for estimating marketing-mix models. More recently, Anand and Lee (2022) utilized a deep learning method, generative adversarial networks (GANs), to generate synthetic data and showed improvement over Schneider et al. (2018) on the privacy-utility tradeoff. The current paper maintains a focus on marketing data but considers reidentification risk arising from potential data linkage. Moreover, we adopt a well-known privacy model, *k*-anonymity, to develop a new protection method for panel data.

In the domain of computer information systems, a large body of work has developed methodologies to protect data privacy; these are known as privacy-preserving data publishing (PPDP), which overlaps with the literature on statistical disclosure limitation. In general, privacy models are classified into two categories that concern two different types of privacy attacks, record linkage and probabilistic attack (Fung et al. 2010). The former type of attack assumes that the intruder knows the QID and uses it to link records to reidentify individuals. Well-known data protection models include *k*-anonymity, *l*-diversity (Machanavajjhala et al. 2006) and *t*-closeness (Li et al. 2007), where the latter two have been developed to address the risk of sensitive-attribute disclosure (Duncan and Lambert 1989; Li and Sarkar 2006), requiring that the values of sensitive attributes in each *k*-anonymized group (i.e., the group of individuals that share the same QID) to be as diverse as possible. The second type of privacy attack, probabilistic attack, does not assume record linkage; instead, it is based on the intruder's prior and posterior probabilistic beliefs about a target after accessing the published data. The most well-known models to prevent probabilistic attacks are based on ϵ-differential privacy (Dwork 2006), which provides a theoretical guarantee that the presence or absence of an individual's record in a database will not substantially affect the outcome of any analysis. As mentioned earlier, the present study concerns privacy attacks through record linkage with the focus on identity disclosure risk.

To prevent record linkage-based privacy attacks, a wide range of methods have been developed under the framework of *k*-anonymity. Common ways to achieve *k*-anonymity are generalization and suppression for which various algorithms have been proposed (Sweeney 2002b; LeFevre et al. 2005, 2006; Zhu et al. 2009). Because the *k*-anonymity problem is *NP*-hard (Meyerson and Williams 2004; Aggarwal et al. 2005), a vast literature mainly focuses on improving computational efficiency and optimality (Aggarwal 2005; Bayardo et al. 2005; LeFevre et al. 2005, 2006; Kenig and Tassa 2012). However, the larger the number of

attributes that constitute a QID, the more difficult it is for generalization to accomplish *k*-anonymity; hence, more observations need to be suppressed. This often results in substantial loss of accuracy in the protected data relative to the original data (Samarati and Sweeney 1998; Aggarwal 2005). Several clustering-based approaches, also known as microaggregation, have been proposed to overcome limitations in generalization and suppression (Domingo-Ferrer and Torra 2005; Nergiz and Clifton 2007). The idea is to partition the data into small subsamples in which individuals are similar in terms of QIDs. Li and Sarkar (2011, 2013) developed tree-based methods to anonymize numeric attributes, for which generalization-based methods can cause significant information loss (Domingo-Ferrer and Torra 2005). Li and Qin (2017) studied medical data and anonymizes text records with a clustering-based approach.

Our approach is different from the literature in terms of both the privacy model and the privacy-enhancing solution. First, as noted, most existing work assumes one record per individual in the target data whereas in panel data each individual has multiple records over time. Any solution to achieve *k*-anonymity must ensure that all QIDs of any given individual are indistinguishable from at least *k* - 1 other individuals. If care is not taken, it is possible that in the protected data *k*-anonymity is satisfied across the same individual's different records but is violated across different individuals. In the household panel data, for example, if household A has two shopping trips with different QIDs, then to achieve 2-anonymity one of the trips may be altered such that the two QIDs become the same in the protected data. Now, even though it appears that 2-anonymity is satisfied in the traditional sense, this QID may still be used to identify household A if no other household has this QID. Kartal and Li (2020) is among the few studies that consider data with multiple records per person. However, they do not assume a longitudinal structure, so that all records of an individual belong to the same *k*-anonymized group. This is fundamentally different from our case, where an individual with multiple records should be in multiple *k*-anonymized groups in order to preserve the heterogeneity and the longitudinal structure of the panel data.

Second, neither generalization nor suppression are suitable approaches to achieve *k*-anonymity for panel data applications. A natural way to generalize household panel data under our setting is to convert the number of units bought of a brand to binary values that indicate only whether the brand was chosen. Such a

generalization suffers from great loss of information as all purchase quantities are discarded, which poses significant limitations for marketing applications. Even with such a generalization, k-anonymity is not guaranteed, even for k=2, as a set of choices can also be unique. Another way to generalize is aggregating some brands. For instance, Lay's Natural and Lay's Wavy are two sub-brands, hence they can be generalized to a single brand Lay's. However, such a transformation reduces the value of the data considerably because the two sub-brands are priced and positioned differently and aggregating them implies that differences as well as competition between them can no longer be examined. The most extreme form of generalization is to aggregate the panel data across households to obtain, say, weekly sales data, whereby all information at the individual level is erased. Although reidentification risk is likely to be fully removed, such aggregation reduces the utility of the protected data drastically. For instance, it is much more challenging to model brand choices accounting for heterogeneity between households, an essential requirement for segmentation and targeting strategies (Bodapati and Gupta 2004; Chen and Yang 2007). Suppression can also result in significant loss of information. As mentioned earlier, to achieve 2-anonymity, suppression involves deleting all records that appear to be unique in the dataset. For larger k, the number of removed observations can be very large. Our empirical study includes special cases of generalization (aggregation) and suppression (record deletion) as benchmarks for comparison and we find that both entail substantial loss of information.

3 Reidentification Risks in Panel Data

3.1 Snowballing Unicity

To provide a visual depiction of how reidentification risk arises, we first return to the household panel data example discussed earlier. Figure 3(a) is a graphical representation of the data shown in Table 1, where the nodes A, B, C and D are the four households, and A1, A2, and A3 represent QIDs associated with A's three shopping trips (see the column labeled "Transaction" in Table 1). Similarly, B1-B3, C1-C3 and D1-D2 are QIDs associated with B, C, and D's shopping trips. Labels on a solid line connecting two households indicate QIDs that are the same. A label on a dashed line indicates a QID that is unique (e.g., A1 in Table 1 and in Figure 3(a)); accordingly, dashed lines do not connect two households. Re-identifiable households are

highlighted in red. To begin, it appears that only household A is re-identifiable, implying that the risk of re-identification is one out of four households, or 25%.

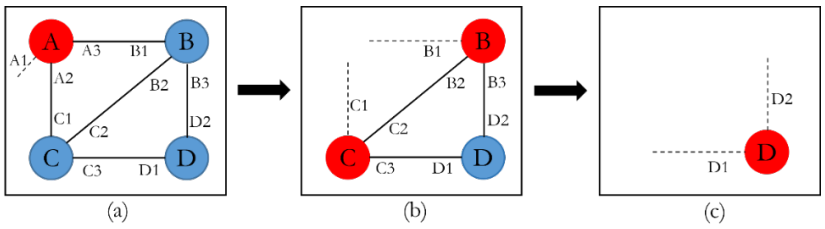

Figure 3: Graphical illustration of sno-unicity based on Table 1. From left (a) to right (c), we illustrate how the reidentification risk is higher than that based on the traditional unicity measure.

Simulating what an intruder would do next if household A is reidentified by a successful match of the unique QID (2, 2, 2), we then remove all records associated with household A, resulting in graph (b) in which households B and C have unique QIDs corresponding to transactions B2 and C1, respectively. Therefore, B and C can be further uniquely reidentified if both B1 and C1 can be matched to external data. After removing all records of households B and C, we find that household D is also re-identifiable as shown in panel (c). As a result, all four households in these panel data can be uniquely re-identified, implying that the reidentification risk is 100% and not 25%, which is what the unicity measure incorrectly showed based on graph (a).

By recursively deleting individuals and re-assessing unicity, we see that the privacy risk is much higher in panel data. Hence the notion of snowballing unicity is naturally suitable for panel data. We formally define snowballing unicity (sno-unicity) below, followed by a procedure that outlines how sno-unicity is computed.

Definition 1 (Snowballing Unicity): Let $\mathcal{Q}_j = \{QID_{j,1}, \dots, QID_{j,t_j}\}$ be the set of all QIDs for panelist j and $\mathbb{Q}^{(1)} = \{\mathcal{Q}_1, \dots, \mathcal{Q}_m\}$ be the set of QIDs for all m panelists. Define $\delta_j^{(1)} = 1$ if there exists a $QID_{j,t}$ such that $QID_{j,t} \notin \mathbb{Q}^{(1)} \backslash \mathcal{Q}_j$ and 0 otherwise. Then we can recursively define $\mathbb{Q}^{(l)} = \mathbb{Q}^{(l-1)} \backslash \mathbb{Q}^*$ for $l = 2, 3, \dots$, where $\mathbb{Q}^* = \{\mathcal{Q}_j \in \mathbb{Q}^{(l-1)} : \delta_j^{(l-1)} = 1\}$, until $\mathbb{Q}^* = \emptyset$. Then the snowballing unicity is defined as

$$SU = \frac{1}{m} \sum_l \sum_j \delta_j^{(l)}.$$

It is not hard to see from Definition 1 that the traditional measure of unicity is calculated as $\frac{1}{m} \sum_j \delta_j^{(1)}$, which is less than or equal to SU. Sno-unicity measures the largest possible proportion of re-identifiable individuals under record linkage, while unicity only reflects that proportion in a single attempt,

ignoring the impact of a reidentified individual on the privacy risk of other individuals. It is worth noting that Definition 1 also defines who is re-identifiable and at which iteration, i.e., panelist j is re-identifiable at iteration l if $\delta_j^{(l)} = 1$. This implies that under an actual privacy attack, *SU* is realized if an intruder has external data where the QID can be exactly matched to those in the target data satisfying $\delta_j^{(l)} = 1$ across all iterations. This can be viewed as a worst-case scenario with respect to the information an intruder may have. In contrast, unicity only defines re-identifiable individuals if $\delta_j^{(1)} = 1$. Thus, for any given external data, the empirical reidentification risk computed based on *SU* is always greater than or equal to that based on unicity.

As discussed earlier, given QID attributes, uniqueness-based measures, both unicity and sno-unicity, are a *property of the dataset itself* that does not depend on the intruder's knowledge. As a result, they are computed purely based on the panel data. The following procedure outlines how sno-unicity is calculated.

Procedure for Calculating snowballing unicity (sno-unicity)

Input: Panel data at the transaction-level
Result: Sno-unicity
Initialization: Construct QID as a separate variable and reshape the panel data to a two-way contingency table with rows being all unique QIDs and columns being unique panelist IDs. Entries in the table are either 1 indicating that a panelist has a specific QID, or 0 if not.
  while *there are rows with row sum equal to 1* **do**
  Decompose the table into two tables:
  Table A: first subset by rows with row sum equal to 1 (this indicates the QIDs are unique). Then subset by columns with column sum greater than or equal to 1 (this indicates the customers who have at least one unique QID). $ncol(A) = \sum_j \delta_j^{(l)}$ is the number of re-identifiable individuals in iteration l;
  Table A^C: The complement of Table A is used for the next iteration if necessary (e.g., evaluated in the "while" condition).
  end
Calculate the proportion of all re-identifiable individuals.

3.2 Individual Reidentification Risk

The above procedure for computing sno-unicity enables us to obtain two byproducts that can be useful in assessing the reidentification risk at the individual level. First, we are able to determine who is re-identifiable and at which iteration. Second, for each re-identifiable individual, we can obtain the number of unique QIDs. For instance, if an individual has 5 different unique QIDs, then this person is easier to re-identify than another individual with only one unique QID. If two individuals have the same number of unique QIDs, but one person is found to be re-identifiable in the first iteration while the other in the second

iteration, then the first person has higher reidentification risk than the second. Based on this intuition, we derive the following probabilistic measure of individual reidentification risk (IR):

Definition 2 (Individual Risk): Let N_j be the total number of distinct QIDs of panelist j, and M_{j,l_j} be the number of those QIDs that are different from all other individuals' QIDs at iteration l_j, the iteration at which panelist j is re-identifiable according to Definition 1. Define $M_{j,l_j} = 0$ if panelist j is not re-identifiable along the snowballing process. Then the reidentification risk of panelist j is defined as

$$IR_j = \frac{M_{j,l_j}}{N_j} \times \overline{IR}_{l_j-1},$$

where $\overline{IR}_l = 1$ for $l = 0$, and $\overline{IR}_l = \frac{\sum_j IR_j \times \delta_j^{(l)}}{\sum_j \delta_j^{(l)}}$ for $l \geq 1$, where $\delta_j^{(l)}$ follows Definition 1.

To better understand Definition 2, we start with a re-identifiable panelist j in the first iteration, that is, $IR_j = M_{j,1}/N_j$. This is simply the proportion of QIDs of panelist j that are different from all other individuals' QIDs, which can be viewed as the likelihood of a unique match for panelist j. Moving forward, the re-identifiable individuals in the l_jth iteration are conditional on those who can be reidentified in the $(l_j - 1)$th iteration. Therefore, multiplying $\overline{IR}_{(l_j-1)}$, the average IR of those who are re-identifiable in the previous iteration, with $M_{j,l_j}/N_j$ results in the reidentification probability for panelist j in the l_jth iteration.

Measurement of individual reidentification risk can provide useful insights about the buying behaviors of panelists who are at higher risk of reidentification. Our empirical application to the IRI data suggests that panelists who are heavier category buyers and less brand loyal are more re-identifiable.

3.3 Additional Comments on Uniqueness-based Reidentification Risk

We conclude this section by making a few additional comments on uniqueness-based measures of reidentification risk such as unicity and sno-unicity. First, the reidentification risk measured by unicity or sno-unicity arises from the privacy attack based on record linkage with deterministic matching. The concept of record linkage originated in data integration applications (Fellegi and Sunter 1996) where the goal is to *increase the chance of correct matching* such that the combined data contains more accurate information. By contrast, in

data privacy applications if a match leads to reidentification then privacy is compromised; as a consequence the goal is to *lower the chance of correct matching* to enhance privacy. Deterministic and probabilistic matching are two common matching methods. Zhu et al. (2015) demonstrated based on extensive simulation studies that deterministic matching works well for high quality data where identifiers can be uniquely matched, while probabilistic matching is preferred for poorer quality data that contain many duplicate or erroneous records due to which deterministic matching is likely to return either multiple records or false matches. Although probabilistic matching is often preferred in modern data warehousing and data integration due to large scale entity matching needs (Dey et al. 1998; Dey 2002; Herzog et al. 2007), the panel data we study in this paper have high levels of unique QIDs, implying that deterministic matching can already lead to very high risk of reidentification. In such a situation, preventing privacy attacks with deterministic matching is an immediate need. We leave the study of reidentification risk under probabilistic matching to future research.

Second, although following convention we use uniqueness-based measures to quantify the risk of reidentification, it is notable that unicity or sno-unicity are not the only measures of privacy risk. Imagine a k-anonymized dataset in which there are no unique QIDs. Then although unicity and sno-unicity are both 0, this does not mean that there is no risk of reidentification. In this case, privacy risk can arise from disclosure of certain sensitive attributes, hence it depends on how the sensitive attributes are distributed in each k-anonymized group. For instance, if all individuals who are included in one k-anonymized group based on their QIDs (demographic information) have heart disease, then the risk of disclosure of this sensitive information is 100% for anyone belong to this group, even though QID-based matching leads to k different people. Privacy models such as l-diversity are designed to protect against the risk of disclosure of sensitive information in such scenarios and we discuss this in the end of Section 4.

4 Protecting Panel Data for k-Anonymity

To reduce the data linkage-based reidentification risk, we propose a new method – graph-based minimum movement k-anonymization (k-MM) – that guarantees k-anonymity while ensuring minimal distortion. We define panel data as satisfying k-anonymity if the QID for every record appears for at least k different panelists. Unlike the conventional definition of k-anonymity that ignores multiple QIDs per

household, the satisfaction of k-anonymity in panel data requires not only that any QID should appear in at least k observations, but also for k different individuals. A self-evident property of k-anonymity is that any column-wise subset of k-anonymized data also satisfies k-anonymity because the subset has fewer attributes in the QID, thereby reducing the amount of information that can be used for linking. In what follows we present the proposed k-MM method that is particularly designed for panel data.

4.1 Graph-based k-Anonymization for Household Panel Data

Consider the illustrative example in Table 1, which has also been graphically represented in Figure 3. Table 1 can be further transformed into the two-way contingency table shown as the matrix $\boldsymbol{A}$ in Table 2, where rows are unique values of QID, and columns are unique panelist IDs. The entries take values 1 or 0, indicating whether or not a specific QID corresponds to a panelist. This same matrix $\boldsymbol{A}$ was constructed in the procedure whereby sno-unicity is computed.

Table 2: Two-way contingency table before (left) and after (right) protection for the example of Table1

Matrix $\boldsymbol{A}$				
Quasi-identifier	Panelist ID			
	A	B	C	D
QID_1 (2, 2, 2)	1	0	0	0
QID_2 (2, 0, 1)	1	0	1	0
QID_3 (2, 2, 0)	1	1	0	0
QID_4 (2, 2, 1)	0	1	1	0
QID_5 (3, 0, 2)	0	1	0	1
QID_6 (3, 1, 0)	0	0	1	1

Matrix $\boldsymbol{B}$				
Quasi-identifier	Panelist ID			
	A	B	C	D
QID_1 (2, 2, 2)	0	0	0	0
QID_2 (2, 0, 1)	1	0	1	0
QID_3 (2, 2, 0)	1	1	0	0
QID_4 (2, 2, 1)	1	1	1	0
QID_5 (3, 0, 2)	0	1	0	1
QID_6 (3, 1, 0)	0	0	1	1

We discussed previously (and also saw in Figure 3) that household A is re-identifiable due to the unique QID associated with transaction A1, after which all the remaining households become re-identifiable sequentially. This can also be seen from Table 2(matrix $\boldsymbol{A}$), where the first row has only one "1" in column A, indicating that QID_1 (2, 2, 2) is uniquely "owned" by household A, hence A is re-identifiable. If household A is reidentified, column A can be deleted; thereafter households B and C become re-identifiable due to uniqueness of QID_3 and QID_2 respectively. Then, deleting column B and C leads household D re-identifiable.

This example illustrates that to achieve k-anonymity, each row of matrix $\boldsymbol{A}$ in Table 2 needs to either have at least k nonzero entries, or all entries should be zero. If k = 2, row QID_1 does not meet this criterion. There are three alternative solutions to achieve 2-anonymity, which are also graphically represented in the three panels of Figure 4: (1) remove all transaction records associated with QID_1, namely, transaction A1 of

household A; see Figure 4 panel (a). (2) add a fake transaction record for households B, C or D that is identical to QID_1, resulting in another 1 in row QID_1. For instance, make up transaction D3 for D so that A1 and D3 have the same *QID;* see panel (b). (3) alter some attributes of QID_1 to make it the same as another *QID,* e.g., QID_4; see panel (c). Corresponding to Figure 4(c), matrix $\boldsymbol{B}$ in Table 2 shows that the protection alters QID_1 to be the same as QID_4, so that 2-anonymity is achieved.

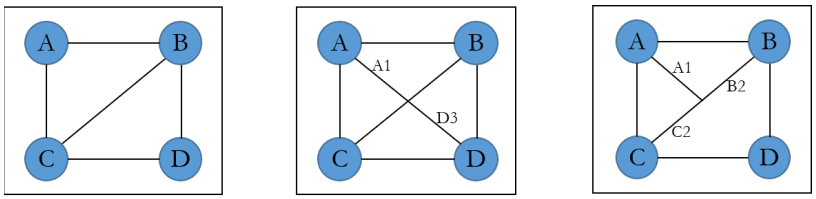

Figure 4: A graphical representation of *k*-anonymization for the household panel data in Table 2. From left to right the three types of solutions are: (a) deletion; (b) addition; (c) altering.

We adopt the altering approach (*i.e.*, Figure 4(c) and Table 2 matrix $\boldsymbol{B}$) for the development of our *k*-MM approach because the other two strategies – deletion and addition – result in larger distortion of the original data. For instance, if 20% of transactions are unique in terms of QIDs, we would either delete or make up this number of transactions in order to achieve 2-anonymity. For larger *k*, this number could be even larger. Based on our empirical analysis of the household panel data, deleting records to achieve *k*-anonymity causes 24% (for k=2) to 55% (for k=7) reduction in total number of observations in the unprotected data, which feels unacceptably large (see Table A10 in Appendix D).

In general, panel data can be considered a pseudo-hypergraph $HG(V, E)$, where V is the set of vertices and E is the set of edges. For example, in Figure 3(a) or Table 2 (matrix $\boldsymbol{A}$), $V=\{A, B, C, D\}$ and $E=\{QID_1, \ldots, QID_6\}$. It can be a hypergraph because some edges may connect to multiple (>2) vertices, such as Figure 4(c) where QID_4 (A1, B2, and C2) connects A, B and C. We call it a pseudo-hypergraph because for unprotected data, some edges do not connect to a second vertex such as A1 in Figure 3(a). After achieving *k*-anonymity, the graph becomes a hypergraph, where each edge connects to at least *k* vertices. We denote *k*-$HG(V^*, E^*)$ the *k*-anonymized pseudo-hypergraph, where the set of vertices $V^* = V$ while the set of hyperedges $E^* \subseteq E$. This is because some QIDs may no longer exist after the alteration (e.g., QID_1 in Table 2 matrix $\boldsymbol{B}$). Because such a graphical representation of the problem and its solution is possible, we call our

approach graph-based *k*-anonymization. We illustrate the idea of minimum movement and formulate our *k*-MM approach in the next subsection. For notational simplicity, we do not use the graph notation.

4.2 Formulation of the *k*-MM Approach

Suppose there are m panelists in the panel data. Denote ID_j the jth individual's ID, $j = 1, \ldots, m$. We first construct the quasi-identifier for each row and let $\mathcal{S}_Q = \{\mathrm{QID}_i : i = 1, \ldots, n\}$ be the set of unique quasi-identifiers among all QIDs. The $n \times m$ matrix $\boldsymbol{A}$, as illustrated in Table 2 (matrix $\boldsymbol{A}$), is then constructed, and we denote its entry at ith row and jth column as a_{ij}. Matrix $\boldsymbol{A}$ is likely a large but sparse matrix, where its binary entries indicate whether or not the quasi-identifier QID_i is associated with panelist j. Clearly, if the sum of row i equals 1, then QID_i only appears once in the data, hence the corresponding panelist is re-identifiable.

To achieve *k*-anonymity, we attempt to find a matrix $\boldsymbol{B} = \{b\}_{i,j}$, as illustrated in Table 2, which is a transformation of matrix $\boldsymbol{A}$ with the same dimensions and the same row and column names, such that the sum of every row is either 0 or larger than or equal to k, and then transform matrix $\boldsymbol{B}$ back to the form of panel data. In Table 2, for example, the transformation from matrix $\boldsymbol{A}$ to $\boldsymbol{B}$ is to change the first column from (1, 1, 1, 0, 0, 0) to (0, 1, 1, 1, 0, 0) and leave all the other columns unchanged. This is equivalent to changing the unique *QID*$_1$ such that it becomes the same as *QID*$_4$. The resulting matrix $\boldsymbol{B}$ satisfies 2-anonymity because the sum of the first row is zero and the sums of all other rows are at least 2, implying that none of the QIDs are unique to any panelist. The row sum of zero in matrix $\boldsymbol{B}$ means that the corresponding QID values no longer exist after protection. Transforming matrix $\boldsymbol{A}$ before protection to $\boldsymbol{B}$ after protection essentially involves moving certain nonzero entries in matrix $\boldsymbol{A}$, which is equivalent to altering the values of corresponding QIDs. However, such data alteration, or equivalently, the movement of nonzero entries in matrix $\boldsymbol{A}$, causes information loss compared to the original data. Therefore, solving for an appropriate matrix $\boldsymbol{B}$ with minimum distortion is the key to our approach.

We represent this as an optimization problem, where a natural objective function is to minimize the total amount of change of the values of QIDs. In other words, if a "one" in a column of matrix $\boldsymbol{A}$ needs to move, it should be moved to the row with the most similar QID. Similarity here is defined using distance

measures of the vector of attributes constituting QIDs. For example, in Table 2 (matrix $\boldsymbol{A}$) $QID_1 = (2, 2, 2)$ is unique, and the closest QID in terms of Euclidean distance is $QID_4 = (2, 2, 1)$; therefore, we change the number of units of Ruffles purchased in transaction A1 from 2 to 1.

As discussed, the data user may prefer to have smaller distortion in certain attributes than others based on their use cases. For instance, household panel data are commonly used to estimate logit brand choice models, which are used to determine equilibrium prices of competing brands. The equilibrium mark-ups of manufacturers in such a setting depend on weekly market shares of brands (Besanko et al. 1998). Therefore, brand manufacturers may prefer to have less distortion in the purchases of larger-share brands so that the optimal mark-ups obtained from the protected data remain close to those from the unprotected data. Our approach is generalizable to different use cases and permits such flexibility by allowing a pre-specified vector of weights on the attributes of the QID to be included in computing the distances. For instance, in the pricing application we can use the market shares of brands in the unprotected data as the weight vector, implying that distortions in purchases of larger-share brands are more costly in computing the distance, hence resulting in less distortion in the transformation from matrix $\boldsymbol{A}$ to $\boldsymbol{B}$.

To formulate such an optimization problem, let $\boldsymbol{Q} = (Q_1, \dots, Q_L)$ be an $n \times L$ matrix where rows are n unique QIDs and columns are the L attributes of QID. We first obtain an $n \times n$ weighted distance matrix among rows of $\boldsymbol{Q}(\boldsymbol{w}) = (Q_1 w_1, \dots, Q_L w_L)$, where $\boldsymbol{w} = (w_1, \dots, w_L)^T$ is a pre-specified weight vector for the L attributes. Denote $\boldsymbol{d}_i(\boldsymbol{w})$ the ith column (or row) of the weighted distance matrix. Define decision vector $\boldsymbol{z}_{ij}$ to be a unit vector of length n if $a_{ij} \neq 0$, and a zero vector if $a_{ij} = 0$. Then the distance of moving a_{ij} to $a_{i'j}$ in matrix $\boldsymbol{A}$ can be represented as $\boldsymbol{z}_{ij}^T \boldsymbol{d}_i(\boldsymbol{w})$, where the i'th element in $\boldsymbol{z}_{ij}$ is 1. Formally, define

$$\boldsymbol{z}_{ij} = \begin{cases} \boldsymbol{i}_n & \text{if } a_{ij} \neq 0 \text{ and it is not moved;} \\ \boldsymbol{i'}_n & \text{if } a_{ij} \neq 0 \text{ and it is moved from } i \text{ to } i', i' \neq i; \\ \boldsymbol{0}_n & \text{if } a_{ij} = 0, \end{cases} \tag{1}$$

where $\boldsymbol{i}_n$ $(\boldsymbol{i'}_n) = (0, \dots, 0, 1, 0, \dots, 0)$ with the ith (i'th) element being 1 and rest 0's. Our objective function is to minimize the total weighted distance caused by the movement of nonzero entries in matrix $\boldsymbol{A}$. Thus, the optimization problem can be formulated as:

$$\min \quad \sum_{j=1}^{m} \sum_{\{i:a_{ij}\neq 0\}} \boldsymbol{z}_{ij}^{T} \boldsymbol{d}_i(\boldsymbol{w}), \tag{2}$$

$$\text{subject to: } \sum_{j=1}^{m} b_{ij} > k-1 \quad \text{or} \quad \sum_{j=1}^{m} b_{ij} = 0 \quad \text{for each row } i, i = 1,\ldots,n, \tag{3}$$

$$b_{ij} = 0 \text{ or } 1, \tag{4}$$

where b_{ij} is the entry of ith row and jth column in matrix $\boldsymbol{B}$.

The objective function (2) minimizes the total weighted distance moved. In general, the value of (2) quantifies the distortion resulting from data protection. Constraints (3) guarantee that the row sum of matrix $\boldsymbol{B}$ is 0 or at least k. Together with constraint (4), k-anonymity is guaranteed for the protected data. Note that the constraint in (4) is needed because otherwise it is possible that some elements in $\sum_i \boldsymbol{z}_{ij}$ are greater than 1, in which case constraint (3) is still satisfied but k-anonymity may not be achieved.

The matrix $\boldsymbol{B}$ can be represented using the decision vector $\boldsymbol{z}_{ij}$. Specifically, $\boldsymbol{b}_{\cdot j} = \sum_i \boldsymbol{z}_{ij}$, hence

$$\boldsymbol{B} = [\textstyle\sum_i \boldsymbol{z}_{i1} \quad \sum_i \boldsymbol{z}_{i2} \quad \cdots \quad \sum_i \boldsymbol{z}_{im}]_{n\times m}. \tag{5}$$

To see (6), let $\boldsymbol{Z}_j = [\boldsymbol{z}_{1j} \quad \boldsymbol{z}_{2j} \quad \cdots \quad \boldsymbol{z}_{nj}]_{n\times n}$, and denote $\boldsymbol{a}_{\cdot j}$ and $\boldsymbol{b}_{\cdot j}$ the jth columns of matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. According to the definition of $\boldsymbol{z}_{ij}$ in (1), it is easy to see that $\boldsymbol{Z}_j$ is a matrix such that $\boldsymbol{Z}_j \boldsymbol{a}_{\cdot j} = \boldsymbol{b}_{\cdot j}$. Since $\boldsymbol{z}_{ij} = \boldsymbol{0}$ if $a_{ij} = 0$, then $\boldsymbol{Z}_j \boldsymbol{a}_{\cdot j} = \sum_i \boldsymbol{z}_{ij} = \boldsymbol{b}_{\cdot j}$.

Since the k-anonymization problem is *NP*-hard, for computational efficiency, we propose the following heuristic by adopting the idea of divide-and-conquer. More specifically, instead of solving the problem for the entire data, we split the matrix $\boldsymbol{A}$ row-wise into several sub-matrices and solve it for each sub-matrix separately. In fact, our graph representation and formulation illustrate the opportunity to use a divide-and-conquer method because constraints (3) and (4) apply to each row of matrix $\boldsymbol{A}$ independently. We conduct extensive simulation studies to demonstrate that the proposed heuristic substantially reduces the computing time at a modest cost of optimality (see Appendix B.1). Our simulation also explores the scale of subproblems for different sizes of the original panel data and shows that the computational cost is low for panel data of moderate to large size (see Appendix B.2). We implement our data protection method in Gurobi (9.1.2), and the details can be found in Appendix A.

A heuristic for the k-MM approach

Input: Panel data at transaction-level

Result: *k*-anonymized household panel data at transaction-level

1. Construct QIDs based on the input data and transform it to matrix $\boldsymbol{A}$. Meanwhile, obtain the matrix of weighted distance between all QIDs.[4]
2. Randomly (or use clustering-based strategy) split matrix $\boldsymbol{A}$ row-wise into q submatrices $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(q)}$ that have similar number of rows.[5]
3. For each submatrix, solve the optimization problem (1)-(4), and obtain the solutions $B^{(1)^*}, \ldots, B^{(q)^*}$ as in (A5, see Appendix A). Then combine them to obtain the solution $\boldsymbol{B}^*$.
4. Transform matrix $\boldsymbol{B}^*$ back to the transaction level data.

The proposed *k*-MM approach can be further extended to achieve *l*-diversity (Machanavajjhala et al. 2006), a privacy model that goes beyond *k*-anonymity. As mentioned, *l*-diversity is intended to protect against (sensitive) attribute disclosure. It requires that the sensitive attribute of individuals in each *k*-anonymized group must take at least *l* different values. The graph representation of the proposed *k*-MM approach (i.e., matrix $\boldsymbol{A}$ and $\boldsymbol{B}$) natually facilitates an additional constraint to accommodate *l*-diversity. To see this, let $\boldsymbol{s} = (S_1, \ldots, S_m)$ be the values of the senstitive attribute of the m individuals with *p* distinct values. Thinking of the matrix $\boldsymbol{B}$ described previously, *l*-diversity can be achieved as follows: for each row of $\boldsymbol{B}$ where the row sum is nonzero, the set $\{S_j : b_{ij} \neq 0\}$ must have at least *l* distinct values. Mathematically, this constraint can be represented as

$$\|\boldsymbol{b}_{i\cdot}\boldsymbol{S}_{dum}\|_0 \geq l \quad \text{or} \quad \|\boldsymbol{b}_{i\cdot}\boldsymbol{S}_{dum}\|_0 = 0 \quad \text{for each row } i, i = 1, \ldots, n, \tag{6}$$

where $\boldsymbol{b}_{i\cdot} = (b_{i1}, b_{i2}, \ldots, b_{im})$ is the ith row of matrix $\boldsymbol{B}$, $\boldsymbol{S}_{dum}$ is the $m \times p$ dummy matrix converted from the sensitive attribute $\boldsymbol{s}$, and $\|\boldsymbol{a}\|_0$ is the cardinality of vector $\boldsymbol{a}$, which is the number of nonzero entries. Constraint (6) can be formulated as a set of linear constraints (shown in Appendix A), so that the entire optimization problem remains a linear program and the proposed heuristic still applies.[6]

[4] A small random variable (e.g., uniform between 0 and 0.1) can be added in calculating the distance matrix to help resolve ties.
[5] In our empirical analysis, we use random split and set q=5 to make computing time practically feasible.
[6] Our empirical results focus on *k*-anonymity only and do not include results for *l*-diversity. The reasons are the following: (1) this paper focuses on protecting against identity disclosure risk while *l*-diversity is intended to protect against attribute disclosure risk, and in our empirical application we do not assume any sensitive attributes; (2) we compare the performance of our approach with benchmark methods that are all intended for *k*-anonymity due to the focus of this study; it is unclear how those benchmark methods can be extended to achieve *l*-diversity. Thus, we only report results of *k*-anonymity for a fair comparison.

5 Empirical Applications

Panel data are procured and analyzed by manufacturers to obtain insights into buyer behaviors and to develop marketing strategies. Protection of the data by alteration will inevitably affect these insights. We discuss the results of application of our protection approach to panel data in two important marketing contexts: household purchases of consumer products, and physican choices of prescription drug treatments. In both settings the data are gathered by market research agencies who provide assurances of privacy to the panelists (households and physicians, respectively). While in this section we discuss in depth the application to household data, for the physician data we only show summary metrics in section 5.5 for reasons of space.

5.1 Reidentification Risk for Salty Snack Category

We analyze IRI panel data (Bronnenberg et al. 2008) on purchasing of salty snacks for the Eau Claire, WI market for the year 2012. The salty snack category is suitable because multi-brand, multi-unit buying on a single shopping trip is rare, making the data more suitable for brand choice modeling. We limit our sample to purchases made in the largest store by category volume. First, we report the reidentification risk as we include a successively increasing number of the largest market-share brands, ranging from 3 to 15. As in our earlier example, the QID consists of the purchase week and the number of units of the brands purchased. We expect that as the number of brands included in the analysis increases, the reidentification risk should grow because more attributes are available to make a QID unique.[7]

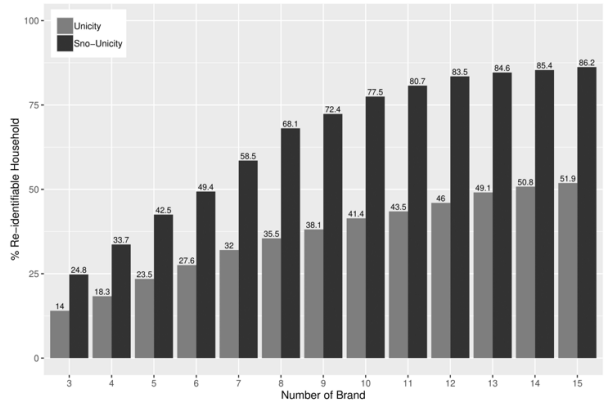

Figure 5: Reidentification risk based on unicity and sno-unicity for different number of brands considered in QID in the salty snacks category in the largest retail store in Eau Claire, WI.

[7] We do not include marketing mix variables such as prices as QIDs because these are homogeneous across shopping trips of different households within a store-week.

Figure 5 shows the reidentification risk measured in two ways: first, using the traditional measure of unicity, and second, using our proposed sno-unicity measure. We see that the difference between unicity and sno-unicity is large regardless of the number of brands considered, indicating that the unicity measure greatly understates the privacy risks in the data. For instance, when the largest 10 brands are included in the sample, unicity reports that 41.4% of households are re-identifiable, whereas the true percentage reported by sno-unicity is alarmingly high at 77.5%. We also observe that as the number of brands increases, the risk of reidentification increases as expected. We find similar patterns for other product categories as shown in Tables A3 and A4 in Appendix D. The reason is that brands being added to the analysis have smaller market shares (see Table A5 in Appendix D), hence they are less likely to be found in QIDs. This implies that less popular brands do not help an intruder identify a large number of panelists, although the few customers who do purchase these brands might be easy to identify.

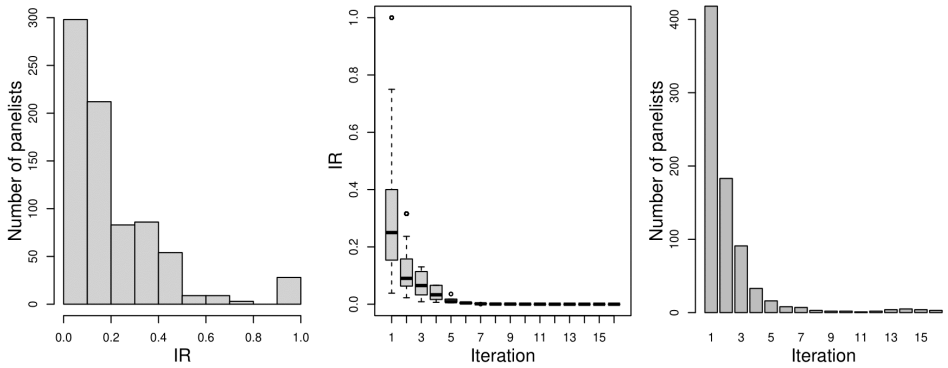

Figure 6: Distribution of individual reidentification risk (IR)

Next, we assess the individual reidentification risk (IR, as in Definition 2) for the case of 10 brands. This sample includes 6,117 shopping trips and 1,009 household panelists. As noted, in this case 77.5% of the 1,009 panelists are reidentifiable. Figure 6 shows, from left to right, the histogram of IR for all re-identifiable panelists; the distribution of IR by iteration l; and the number of re-identifiable panelists in each iteration. We find considerable heterogeneity in the IR across panelists. To assess whether high reidentification risk is associated with observable purchasing behaviors, we construct three variables for each panelist: average number of units purchased per transaction, number of brands ever bought, and total dollar spend across trips. We use these variables as predictors in a logistic model to classify panelists into high and low IR split by median IR. Coefficient estimates are shown in Appendix D (Table A6). We find that all three predictors are

statistically significant, and the signs are intuitive. Heavier buyer (more total spend and more units bought per transaction) and those with less brand loyalty (more brands ever bought) are at greater risk of reidentification.

The following subsections assess the loss of information due to protection of the household panelists using our proposed *k*-MM approach, for different levels of *k*. To assess the loss of data utility, we consider three analyses. First, we assess the changes in the original data and quantify the overall distortion using two metrics. Second, we measure the distortion in three commonly used marketing metrics. Third, we measure the changes in the parameter estimates of a brand choice model applied to the data. Our results are based on protected data with equally-weighted QID attributes. In Section 5.6 we demonstrate the use of unequal weights in the context of determining equilibrium prices.

5.2 Overall Data Distortion

We consider two measures of the overall distortion of the protected relative to the unprotected (transaction-level) data. These two measures are based on a cell-by-cell comparison of the two datasets. To define notation, say each of the unprotected and protected transaction-level datasets have T observations and L columns (L attributes in QID), and let w_j be the weight for attribute l. Let y_{ij} be the entry in the ith row and jth column of the unprotected data, and y'_{ij} be the corresponding entry in the protected data.

Changed cells % is defined as

$$\text{Prop}\{y_{ij} \neq y'_{ij}\} = \frac{1}{T \times L} \sum_{i=1}^{T} \sum_{j=1}^{L} \mathbb{I}(y_{ij} \neq y'_{ij}) \times 100\%$$

Mean squared deviation (MSD) is defined as

$$\text{MSD} = \frac{1}{T \times L} \sum_{i=1}^{T} \sum_{j=1}^{L} w_j (y_{ij} - y'_{ij})^2.$$

The second measure, mean squared deviation, is equivalent to (2) when Euclidean distance is applied.

Table 3 (Panel A) summarizes the information loss in our *k*-MM anonymized data based on these two metrics. For each *k*, the reported results are averages over 100 replicates due to the randomness introduced in the proposed heuristic. These metrics reveal that as expected, there is larger information loss with increasing *k*. This is a demonstration of the classic R-U tradeoff curve between privacy (risk) and accuracy (utility) (Duncan and Stokes 2004; Reiter 2005). However, even when *k*=7, implying that any

panelist is not different from at least six other panelists in the data, the percentage of cells changed in the original data is very small (3.337%).

Table 3: Information loss due to protection for different levels of k-anonymity.

	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
Panel A. Overall information loss in transaction-level data						
Changed Cells (%)	1.17	1.687	2.185	2.624	2.984	3.337
MSD	0.094	0.11	0.128	0.14	0.149	0.157
Panel B. MAPD (%) in brand metrics						
Market share	0.442	0.709	0.983	0.965	1.158	1.229
Share of Category requirements	0.664	0.529	0.681	0.733	0.772	0.857
Brand Switching probability	1.021	0.993	1.324	1.087	1.285	1.233

5.3 Distortion in Marketing Metrics

Next, we consider the distortion in three marketing metrics that are commonly computed using household panel data. The degree of distortion for each metric is measured as the mean absolute percentage deviation (MAPD)[8] between the metrics based on unprotected and protected data. **Brand market share** is defined as the number of choices of each brand divided by the total number of choices of all studied brands in the sample, expressed as a percentage. **Share of category requirements** (SCR) is commonly used as a measure of brand loyalty (Bowman and Narayandas 2001). For each brand j, the SCR is measured at the panelist level as the number of purchases of brand j divided by the total number of purchases in the category, expressed as a percentage and averaged across panelists. **Brand switching** is a vector of length J, computed from the $J \times J$ brand switching matrix $\boldsymbol{S}$. The entry in the jth row and j'th column of $\boldsymbol{S}$, denoted as $s_{jj\prime}$, is the count of pairs of consecutive trips within a panelist wherein brand j was chosen followed by brand j', summed across panelists. Then brand switching for brand j is defined as $\left(1 - \frac{s_{jj}}{\sum_{j'=1}^{J} s_{jj'}}\right) * 100\%$.

As a result of the small distortions in the data discussed previously, we expect to see, and find in Table 3 (Panel B), that the distortion in key marketing metrics is very small as well. For instance, with k=7 the

[8] MAPD is defined as $\frac{1}{J}\sum_{j=1}^{J}\left|\frac{x\prime_j - x_j}{x_j}\right| \times 100\%$, where x_j and x'_j are statistics (in general) of brand j based on unprotected and protected data, respectively.

average absolute distortion in market share of each brand is 1.229% (this is not in percentage points, but a percentage of the true market share).

5.4 Distortion in Parameters of Brand Choice Model

A common application of panel data is brand choice modeling, which reveals intrinsic brand preferences, responsiveness to marketing mix variables, as well as heterogeneity in these parameters across panelists. Here we use the widely employed Hierarchical Bayesian (HB) random effects multinomial logit model (Allenby and Rossi, 1998). Details of the model are shown in Appendix C. We enhance the transaction-level brand choices in the unprotected data with information on the following marketing mix variables for each of the ten brands: price per 16-ounce (standardized) unit, and binary indicators of Promotion, in-Store Display, and Feature Advertising.[9] Table 4 shows the market share and descriptive statistics of the marketing mix variables of each of the ten brands.

Table 4: Share of choices and summary statistics of marketing mix variables for largest selling ten brands.

			Proportion of shopping trips[10]		
Brand	Share of choices	Mean Retail Price ($ per 16-ounces)	Promotion	In-Store Display	Feature Advertising
Lays Natural	0.193	4.867	0.461	0.690	0.278
Old Dutch	0.169	3.801	0.734	0.130	0.053
Cheetos	0.110	4.890	0.371	0.137	0.170
Barrel O Fun	0.094	3.545	0.161	0.064	0.029
Sunchips	0.089	5.168	0.812	0.594	0.138
Lays	0.083	5.196	0.404	0.042	0.246
Tostitos Natural	0.077	3.819	0.863	0.530	0.217
Doritos	0.065	4.368	0.325	0.202	0.068
Wavy Lays	0.062	4.759	0.701	0.156	0.284
Old Dutch Ripples	0.059	5.553	0.737	0.419	0.139

We show in Table 5 parameter estimates based on the unprotected data and k-anonymized data, for each of k = 2,...,7. For each k, we estimate models on 100 replicates of protected data in order to capture the variability due to the randomness included in the proposed heuristic. For the unprotected data we show the mean coefficient estimates over all households based on the last 100 MCMC draws from the posterior

[9] Details of data preparation can be found in Web Appendix C.

[10] For some brands the proportion of promotion and in-store display may appear to be very high to those who are familiar with point-of-sale data for frequently purchased goods. We also computed these proportions in a few other categories and found that the numbers were particularly high for salty snacks, suggesting this may be a characteristic of this category.

distribution. For the protected data, for each k, we show the average across the 100 replications of the mean coefficient estimates based on the last 100 MCMC draws from the posterior distribution. Comparing the unprotected and protected coefficient estimates, none of the signs change relative to the unprotected data, and the magnitudes are very similar. We also present the McFadden R^2 values (again, for the protected data these are averages across the 100 replications), which also change negligibly. To quantify the loss of utility due to protection, we show in the last row of Table 5 the MAPD of the coefficient estimates for the four marketing mix variables. We can see that the overall trend is for MAPD to increase with k, again indicating the tradeoff between reidentification risk and data utility. However, the magnitude of the deviation between estimates based on the unprotected data versus protected data is small, ranging from 6.17% for k=2 to 8.75% for k=7, where k=7 implies a higher level of privacy.

Table 5: Mean values of parameter estimates from Hierarchical Bayesian random effects logit model.[11]

	Unprotected	k-anonymized Data					
Marketing mix variables:	Data	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
logPrice	-1.453	-1.351	-1.348	-1.393	-1.369	-1.288	-1.292
Promotion	0.986	1.101	1.095	1.123	1.122	1.114	1.157
Display	3.350	3.251	3.252	3.264	3.275	3.283	3.246
Feature	1.675	1.727	1.740	1.675	1.651	1.656	1.617
McFadden R^2	0.609	0.618	0.618	0.618	0.615	0.615	0.611
MAPD of marketing mix coefficient estimates (%)	-	6.171	6.274	5.164	5.812	6.877	8.753

Finally, Figure 7 shows the estimated posterior distributions of the coefficients for marketing mix variables based on unprotected and protected data for each k. Results for the protected data for each k are based on one randomly chosen replicate out of 100. We see that for each marketing mix variable the distributions based on protected and unprotected data are similar, again indicating that the protected data preserve the utility of the original data with regard to learning about heterogeneity in consumers' responsiveness to marketing activities. Importantly, as discussed these heterogeneity distributions cannot be estimated reliably if aggregation is employed to protect the data.

[11] Brand-specific constants are shown in Web Appendix D, Table A8, for reasons of space. Posterior standard deviations of the coefficient estimates are reported in Table A9. The results for protected data are averages over 100 replicates.

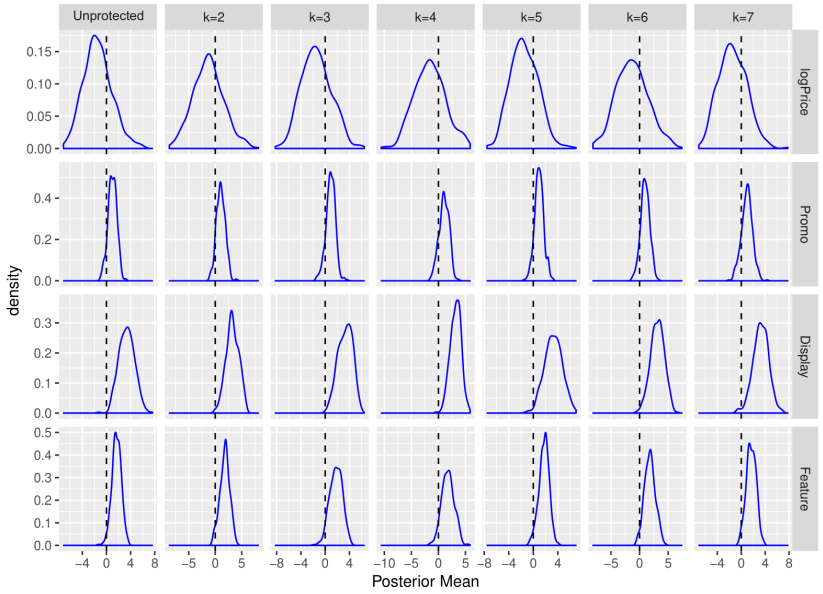

Figure 7: Distribution of posterior means for individual coefficient estimates of marketing mix variables.

5.5 Comparison with Other Data Protection Approaches

We compare our proposed approach to several alternative data protection methods, including (1) aggregation, (2) clustering, (3) record deletion, (4) random swapping, and (5) noise addition based on an ϵ-differential privacy (ϵ-DP) model. Table 6 briefly describes each of these approaches.

Table 6: Benchmark data protection approaches

Privacy protection method	Protection mechanism	Privacy achieved
Aggregation	Aggregate purchases across panelists in each week, so that the protected data are weekly sales data. Aggregation involves summing individual choices by brand and averaging market-mix variables, within each week.	N/A
Clustering	For each week, find certain number of clusters based on the values of QIDs. Then aggregate purchases of all panelists within each cluster. A detailed implementation is outlined in Appendix E.	k-anonymity
Record deletion	Delete the purchases that have up to k-2 duplicates across different panelists. See Figure 4(a) for illustration.	k-anonymity
Random swapping	Randomly choose α% of observations and swap their purchases.	α% alteration
Noise addition	Following Schneider et al. (2017), synthetic brand choices are generated according to different level of privacy that is controlled by ϵ, $\epsilon > 0$; the smaller is ϵ the stronger the protection. Detailed implementation is outlined in Appendix E.	Synthetic data

Among these benchmark approaches, both clustering and record deletion can achieve k-anonymity, hence results from these approaches can be directly compared with our proposed method. Figure 8 shows that for the same privacy level (k), both clustering and simple deletion lead to substantially higher information

loss than our proposed method. The compared metrics include the MAPD of the coefficient estimates for the brand choice model (e.g., last row of Table 5) and the descriptive statistics reported in Table 3. We do not report changed cells and MSD for the record deletion method because this protection mechanism reduces the amount of data by construction, so these two metrics are not defined. It is notable that a disadvantage of clustering is that the choice of the number of clusters to achieve a particular k is not trivial. Sometimes k-anonymity may not be achieved if there is a small number of unique customers in a given week. For example, if there are only 2 unique customers who made purchases during a particular week, then clustering can achieve at most 2-anonymity by altering the purchases of the 2 customers to be the same.

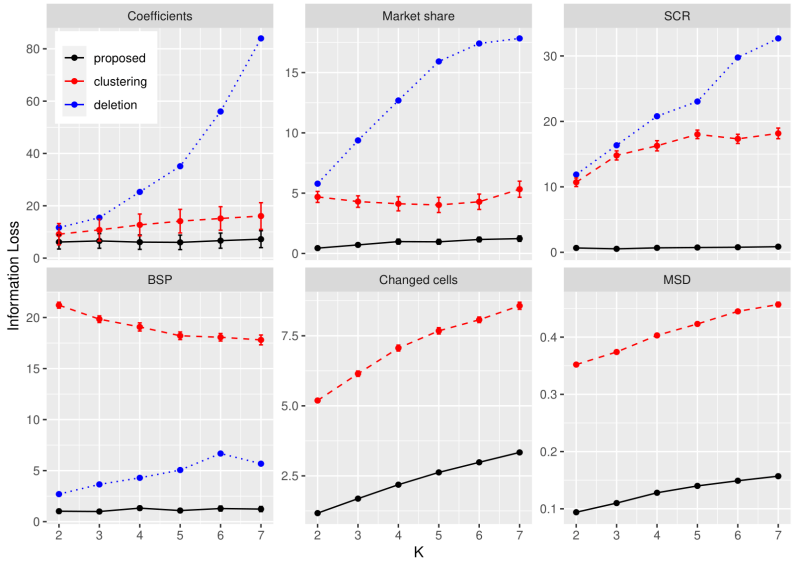

Figure 8: Comparison of information loss between proposed method and two benchmark methods – clustering and deletion – in the salty snacks data. For coefficients, market share, SCR (share of category requirements) and BSP (brand switching probability), the measure of information loss on the vertical axis is MAPD (%). Percentage of changed cells and MSD (mean squared deviation) are measured based on transaction-level data. X-axis shows levels of k for k-anonymity. Error bars show the standard deviation obtained from 100 replicates.[12]

The other three benchmark approaches – aggregation, random swapping, and noise addition – cannot achieve the protection goal of k-anonymity, because they do not reduce the uniqueness of QIDs. As a result, we cannot directly compare the information loss between these methods and our proposed method at particular privacy levels. Instead, in Table 7 we average the information loss of our proposed method across k=2,…,7, and compare it to the information loss incurred at a moderate privacy level for the random

[12] Record deletion does not involve randomness so standard deviation is not computed.

swapping and the noise addition approaches. To illustrate we choose $\alpha = 20$ (20% random swapping) and $\epsilon = 2$, meaning that the odds of any individual being in the data is bounded above by $e^2 (\approx 7.4)$ according to the definition of ϵ-differential privacy[13] (Dwork 2006). Table 7 shows that the proposed method results in the least information loss in every metric. Note that we do not include results from aggregation as this approach completely removes all consumer-level information; as a result, only one comparison metric can be computed – the MAPD of coefficients – which at 55.1% is much larger than the proposed method.[14]

Table 7: Comparison of information loss between proposed method (average across k = 2, …, 7) and 20% random swapping and noise addition ($\epsilon = 2$). Compared metrics are the same as those in Figure 7.

	Proposed (avg. k=2,…,7)	20% random swapping	Noise addition ($\epsilon = 2$)
Coefficients*	6.461	9.326	12.650
Market share*	0.914	1.490	21.555
SCR*	0.706	7.263	13.731
BSP*	1.157	4.564	11.154
Changed cells (%)	2.331	3.720	11.568
Mean Squared Deviation	0.130	0.474	0.843

* The reported measures are MAPD%

5.6 Weighted Cost of Distortion in Determining Equilibrium Mark-ups

We demonstrate that our data protection approach flexibly allows the user to specify different degrees of distortion in different QID attributes. Brand choice models have been used to determine equilibrium wholesale prices in oligopolistic markets. The optimal mark-up on cost for manufacturer of brand j under the assumption of vertical Nash competition between manufacturers selling through a common retailer is given by $\frac{1}{\alpha(1-share_j)}$, where α is the estimated price coefficient from a multinomial logit brand choice model and $share_j$ is the estimated market share of brand j (see equation 7a in Besanko et al. 1998). A user interested primarily in determining equilibrium prices may choose weights that result in smaller distortion in market shares of larger-share brands while allowing greater distortion in market shares of smaller-share brands.

[13] In practice, ϵ ranges from 0.1 to 4.6, which result in the upper bound of odds being 1.1 and 100, respectively.
[14] We apply the traditional logit model to estimate aggregate-level (i.e., homogeneous) parameters, while recognizing that studies such as Berry, Levinsohn and Pakes (1995), Besanko, Dube and Gupta (2003), Musalem et al. (2009) and others use aggregate data to estimate heterogeneous (individual-level) parameters for the brand choice model.

Using the household panel data for salty snacks, we re-run the protection with unequal weights for each of the ten brands. Specifically, we use brands' market shares calculated based on the unprotected data as the weight vector and compute the weighted distance matrix as outlined in Step 1 of our heuristic, while all other steps remain the same. Figure 9 depicts the MAPD (over 100 replicates) of total sales for the three (out of 10) brands that have largest (Lays Natural), median (Wavy Lays), and smallest (Sunchips) market shares. Comparing the unweighted (left panel) and weighted (right panel) results, we see that the largest brand suffers smaller distortion under weighting, the smallest brand suffers larger distortion, while the median brand is relatively unaffected. These findings are as expected based on the formulation of our *k*-MM approach and provide assurance that the proposed protection procedure can be used to differentially distort QID attributes as desired by the data user. Additional simulation results also confirm these findings (see Appendix B.3).

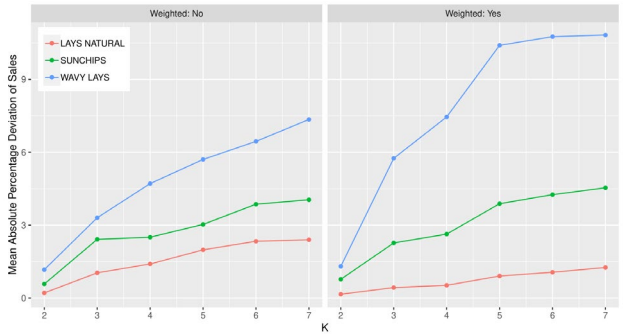

Figure 9: Distortion of sales measured by MAPD of the largest (Lays Natural), median (Sunchips) and smallest (Wavy Lays) market share brands under unweighted (left panel) and weighted (right panel) protection.

5.7 Summary of Results for Physician Prescribing Data

To further demonstrate the usefulness and performance of our proposed k-MM data protection method, we apply it to prescribing behavior of physicians in the US in the statin category. The data were collected by a market research firm from a physician panel that is a representative sample of the physician universe. Our data include 14,995 prescriptions written over a 24-month period by 448 physicians for three major statin brands: Lipitor (produced by Pfizer), Zocor (Merck), and Crestor (AstraZeneca), and an alternative prescription option called "nondrug treatment". The major marketing activity of interest is detailing, or salesperson visits to physicians. A detailed description of the dataset is provided in Liu et al. (2016); we show descriptive statistics in Table A11 in Appendix F. We apply the Hierarchical Bayes random

effects multinomial logit model to the prescription choices of physicians. Coefficient estimates are shown in Table A12 in Appendix F.

Analogous to Figure 8 and Table 7, we compare the information loss due to the proposed protection method and benchmark methods in Figure 10 and Table 8. Again, we find that the proposed method leads to smaller information loss than all benchmark methods, except for small MAPDs of market share.

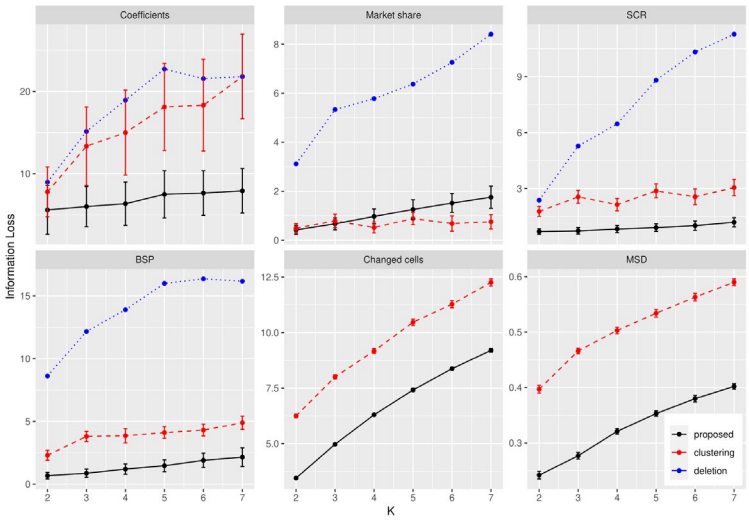

Figure 10: Comparison of information loss between proposed and two benchmark methods – clustering and deletion – for physician prescribing data. For coefficients, market share, SCR (share of category requirements) and BSP (brand switching probability), the information losses are MAPDs (%). Changed cells and MSD (mean squared deviation) are measured based on the transaction-level data. X-axis shows levels of k for k-anonymity. The error bars show the standard error obtained from 100 replicates.

Table 8: Comparison of information loss in physician prescribing data between proposed method (average across $k = 2, \ldots, 7$) and 20% random swapping, noise addition ($\epsilon = 2$), and aggregation. Compared metrics are the same as those in Figure 8.

	Proposed (avg. k=2,…,7)	20% random swapping	Random noise ($\epsilon = 2$)
Coefficients*	6.841	18.312	36.718
Market share*	1.099	1.546	8.373
SCR*	0.892	3.592	10.443
BSP*	1.377	3.514	3.979
Changed cells (%)	6.619	8.674	16.633
Mean Squared Deviation	0.329	0.632	0.886

* The reported measures are MAPD%

6 Conclusion

Our investigation revealed a high risk of reidentification of panelists in two commonly used market research datasets: household panel data gathered by IRI and AC Nielsen, and physician prescription writing behaviors. We have also shown that the traditional measure of reidentification risk – unicity – can

significantly understate the risk because it does not account for the longitudinal nature of the data. We proposed a measure of reidentification risk – sno-unicity – that is better suited to the longitudinal nature of panel data. Finally, we have proposed a solution to protect the data via alteration that guarantees *k*-anonymity and have demonstrated that in both empirical applications, the utility of the data for typical uses by marketing managers is reduced only to a modest degree. Our empirical results also show that compared to several alternative protection methods, the loss of utility resulting from our approach is much smaller.

Our proposed *k*-MM approach has several advantages. First, unlike commonly used generalization and suppression-based methods, our approach neither requires a predefined domain generalization hierarchy, which can be subjective and arbitrary, nor deletes any observations. Second, our approach integrates flexibilities that allow users to choose different levels of distortion for different attributes. Third, the graph representation of our approach naturally facilitates an additional constraint to achieve *l*-diversity, a privacy model that is intended to protect against sensitive attribute disclosure. Finally, and most importantly, the essence of our approach is a linear program that minimizes the *cellwise* distortion between the protected data and the original data. Therefore, a high level of utility in the original data is preserved for any analysis in general, not limited to a specific application. We should also note that our method can be applied not only to numerical but also categorical variables as long as a distance metric is predefined for the vector of QIDs.

Our findings are important because typical approaches to protecting data against reidentification are known to be inadequate to achieve the dual objectives of increasing data privacy and maintaining data utility. These include access-control through data security protocols, whose vulnerability is demonstrated by thousands of data breaches each year (Verizon 2019). Further, several studies have demonstrated that deidentifying data provides insufficient protection against linkage attacks (Narayanan and Shmatikov 2008; Benitez and Malin 2010). This implies that even though data providers may be taking all reasonable precautions to protect their panelists, including those required by law, the uniqueness of purchasing patterns creates a privacy vulnerability. Although much of the data involved in these settings may not be considered sensitive, the reputational damage to both the data-collecting organization and the individual if any sensitive information is leaked is potentially high.

Although our empirical applications focused on two widely used data sources, the method we have proposed can be applied in many settings in which individual-level data are gathered via panels, and where privacy concerns arise. Examples include Sensor Tower's large panel of mobile app users around the world (Rogers 2021), Comscore's panel of individuals who report web browsing, and Nielsen's TV and Radio panels. In each of these cases, the uniqueness of reported behaviors creates the possibility of re-identification of the panelist by an intruder with external data. In addition to the possibility that some data of identified panelists may be sensitive, an additional risk in market research settings is that identified panelists could be incentivized by motivated intruders to record biased data. We recommend that organizations that collect and share purchasing data at the individual-level should consider data protection solutions like ours, in addition to routine measures such as removal of direct demographic identifiers.

Next, we briefly outline some limitations of our work and directions for future research. While panel data gathered via market research remain crucially important in many industries, growing privacy regulations are forcing organizations to seek alternative forms of data. Further, as discussed, our study follows the conventional setting in which reidentification occurs through deterministic data linkage. Reidentification, however, can also be done through probabilistic linkage, which is a practically useful concept in database management, but requires a separate and comprehensive study in the data privacy context. Finally, while we examined the numerical properties of our proposed approach through simulation studies, future work should examine its theoretical properties, such as the rate of convergence and computational complexity.

References

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases,* VLDB '05, 901–909.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., & Zhu, A. (2005). Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT).*

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1-2):57–78.

Anand, P., & Lee, C. (2022). Using Deep Learning to Overcome Privacy and Scalability Issues in Customer Data Transfer. *Marketing Science.*

Bayardo, R. J., & Agrawal, R. (2005, April). Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)* (pp. 217-228).

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, 17(2), 169-177.

Berry, S., Levinsohn J., and Pakes A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63 (4), 841-890.

Besanko, D., Dube JP, and Gupta, S. (2003). Competitive price discrimination in a vertical channel using aggregate retail data. *Management Science*, 49(9), 1121-1138.

Besanko, D., Gupta, S., & Jain, D. (1998). Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Science*, 44(11-part-1), 1533-1547.

Bodapati, Anand V. and Sachin Gupta (2004), "The Recoverability of Segmentation Structure from Store-Level Aggregate Data," *Journal of Marketing Research*, 41 (3), 351–64.

Bowman, D., & Narayandas, D. (2001). Managing Customer-Initiated Contacts with Manufacturers: The Impact on Share of Category Requirements and Word-of-Mouth Behavior. *Journal of Marketing Research*, 38(3), 281-297.

Bronnenberg, Bart J., Michael W. Kruger, and Carl F. Mela. (2008). "Database Paper —The IRI Marketing Data Set." *Marketing Science*, 27 (4):745–48.

Bruno, H. A., Cebollada, J., & Chintagunta, P. K. (2018). Targeting Mr. or Mrs. Smith: Modeling and leveraging intrahousehold heterogeneity in brand choice behavior. *Marketing Science*, 37(4), 631-648.

Bucklin, R. E. and Gupta, S. (1999). Commercial use of upc scanner data: Industry and academic perspectives. *Marketing Science*, 18(3):247–273.

Chen, Yuxin and Sha Yang (2007), "Estimating Disaggregate Models using Aggregate Data through Augmentation of Individual Choice," *Journal of Marketing Research*, 44 (4), 613–21.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75(370), 377-385.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 1376.

De Montjoye, Y. A., Radaelli, L., Singh, V. K., et al. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539.

Dey, D., Sarkar, S., & De, P. (1998). A probabilistic decision model for entity matching in heterogeneous databases. *Management Science*, *44*(10), 1379-1395.

Dey, D. (2003). Record matching in data warehouses: a decision model for data consolidation. *Operations Research*, 51(2), 240-254.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2), 195-212.

Duncan, G. T., & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393), 10-18.

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2), 207-217.

Duncan, G. T. and Stokes, S. L. (2004). Disclosure risk vs. data utility: The RU confidentiality map as applied to topcoding. *Chance*, 17(3), 16-20.

Dwork, Cynthia (2006), "Differential Privacy," 33rd International Colloquium on *Automata, Languages and Programming*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, eds., Springer Berlin Heidelberg, 1-12.

El Emam, K. and Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Ferrell, O. C. (2017). Broadening marketing's contribution to data privacy. *Journal of the Academy of Marketing Science*, 45(2), 160-163.

Finck, M. and Pallas, K. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR, *International Data Privacy Law*, 10(1):11–36.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53.

Gelman, A., & Park, D. K. (2009). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, 63(1), 1-8.

Goldfarb, A., & Tucker, C. (2012). Shifts in privacy concerns. *American Economic Review*, 102(3), 349-53.

Hern, A. (2014, June 27). New York taxi details can be extracted from anonymised data, researchers say. The Guardian. https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

Kappe, E., & Stremersch, S. (2016). Drug detailing and doctors' prescription decisions: the role of information content in the face of competitive entry. *Marketing Science*, 35(6), 915-933.

Kartal, H. B., & Li, X. B. (2020). Protecting Privacy When Sharing and Releasing Data with Multiple Records per Person. *Journal of the Association for Information Systems*, *21*(6), 8.

Kenig, B. and Tassa, T. (2012). A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, 25(1):134–168.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, 9, 313-313.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005). Incognito: Efficient fulldomain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, 49–60.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)* (pp. 25-25). IEEE.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, 106–115.

Li, X. B. & Qin, J. (2017). Anonymizing and Sharing Medical Text Records. *Information Systems Research*, 28 (2), s. 332–352.

Li, X. B., & Sarkar, S. (2006). Privacy protection in data mining: A perturbation approach for categorical data. *Information Systems Research*, *17*(3), 254-270.

Li, X. B., & Sarkar, S. (2011). Protecting privacy against record linkage disclosure: A bounded swapping approach for numeric data. *Information Systems Research*, 22(4), 774-789.

Li, X. B., & Sarkar, S. (2013). Class-restricted clustering and microperturbation for data privacy. *Management Science*, 59(4), 796-812.

Liu, Q., Gupta, S., Venkataraman, S., & Liu, H. (2016). An empirical model of drug detailing: Dynamic competition and policy implications. *Management Science*, 62(8), 2321-2340.

Lipsman, A., Mudd, G., Rich, M., & Bruich, S. (2012). The power of "like": How brands reach (and influence) fans through social-media marketing. *Journal of Advertising research*, *52*(1), 40-52.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, 24–24. IEEE.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008, April). Privacy: Theory meets practice on the map. In *2008 IEEE 24th international conference on data engineering* (pp. 277-286). IEEE.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4), 336-355.

Manchanda, P., Rossi, P. E., & Chintagunta, P. K. (2004). Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, 41(4), 467-478.

Martin, K. D. and Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, 45(2):135–155.

Matthews, G. J. and O. Harel (2011). "Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy." *Statistics Surveys*, 5: 1-29.

Meyerson, A. and Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM.

Musalem, A., Bradlow, E. T., & Raju, J. S. (2008). Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *Journal of Marketing Research*, 45(6), 715-730.

Musalem, A., Bradlow, E. T., & Raju, J. S. (2009). Bayesian estimation of random-coefficients choice models using aggregate data. *Journal of Applied Econometrics*, *24*(3), 490-516.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125.

Reiter, J. P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112.

Rogers, Bruce. "Sensor Tower Builds the 'Nielsen' of the App World." Forbes, Forbes Magazine, 12 Apr. 2021, https://www.forbes.com/sites/brucerogers/2021/04/09/sensor-tower-builds-the-nielsen-of-the-app-world/?sh=4c92f2472272.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, 9(2), 461-468.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International.*

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., and Yu, Y. (2017). Protecting customer privacy when marketing with second-party data. *International Journal of Research in Marketing*, 34(3):593–603.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., and Yu, Y. (2018). A flexible method for protecting marketing data: An application to point-of-sale data. *Marketing Science*, 37(1):153–171.

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS quarterly*, 167-196.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS quarterly*, 989-1015.

Sweeney, L. (2000). Uniqueness of simple demographics in the us population. Technical report, Carnegie Mellon University.

Sweeney, L. (2002a). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Sweeney, L. (2002b). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(05):571–588.

Tucker, C. E. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, 51(5), 546-562.

Verizon. (2019). 2019 Data Breach Investigations Report. https://enterprise.verizon.com/resources/executivebriefs/2019-dbir-executive-brief.pdf.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6), 97-121.

Wieringa, J., Kannan, P. K., Ma, X., Reutterer, T., Risselada, H., & Skiera, B. (2021). Data analytics in a privacy-concerned world. *Journal of Business Research*, 122, 915-925.

Zhu, D., Li, X. B., & Wu, S. (2009). Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. *Decision Support Systems*, 48(1), 133-140.

Zhu, Y., Matsuyama, Y., Ohashi, Y., & Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, 56, 80-86.

Online Appendix for "Reidentification Risk in Panel Data: Protecting for *k*-Anonymity"

August 29, 2022

A. Implementation of *k*-MM approach

Let s_j be the jth column sum of matrix $\boldsymbol{A}_{n\times m}$, and $s = \sum_{j=1}^{m} s_j = \sum_{i,j} \mathbb{I}(a_{ij} \neq 0)$, which is the total number of nonzero entries in matrix $\boldsymbol{A}$. According to (1), there are in total s non-zero decision vectors , i.e. $\{\boldsymbol{z}_{ij}: a_{ij} \neq 0\}$, for which we need to solve. We define a single decision vector $\boldsymbol{z}$ of length $n \times s$, which is a long vector of the s decision vectors $\boldsymbol{z}_{ij}$. Since the constraint (3) has an "or" condition, we need another decision vector of length n, denoting $\boldsymbol{v}$. Then, the constraints (3) and (4) can be re-written as below:

$$C_1\boldsymbol{z} - q\mathbf{I}_n\boldsymbol{v} \leq \mathbf{0}; \tag{A1}$$

$$C_1\boldsymbol{z} - k\mathbf{I}_n\boldsymbol{v} \geq \mathbf{0}; \tag{A2}$$

$$C_2\boldsymbol{z} + \mathbf{0}\boldsymbol{v} = \mathbf{1}; \tag{A3}$$

$$C_3\boldsymbol{z} + \mathbf{0}\boldsymbol{v} \leq \mathbf{1}; \tag{A4}$$

$$(\boldsymbol{z}, \boldsymbol{v}) \text{ are binary.}$$

Here $C_1 = \mathbf{1}_s \otimes \mathbf{I}_n$, is an n by ns matrix, $C_2 = \mathbf{I}_s \otimes \mathbf{1}_n$, is an s by ns matrix, and $C_3 = \mathbf{I}_m \otimes \{\mathbf{1}_{s_j} \otimes \mathbf{I}_n\}_{j=1}^{m}$, is an nm by ns matrix, where $\otimes$ denotes the kronecker product operator. The constant q is an

arbitrary large number such that (A1) and (A2) together represent the constraint (3). More specifically, for those rows of matrix $\boldsymbol{B}$ where $v = 0$, the sum is zero, and for $v = 1$ the sum is greater than or equal to k. The constraint (A3) ensures that each $\boldsymbol{z}_{ij}$ where $\{i: a_{ij} \neq 0\}$, is a unit vector. The inequality (A4) ensures that no element in the matrix $\boldsymbol{B}$ can exceed 1, which is equivalent to constraint (4).

Note that it is possible that an individual may have multiple quasi-identifiers that are identical. This results in a non-binary matrix $\boldsymbol{A}$. In this case, the matrix $\boldsymbol{B}$ as defined in (5) may not be the solution we need. However, since $\boldsymbol{Z}_j \boldsymbol{a}_{\cdot j} = \sum_i \boldsymbol{z}_{ij} = (b_{1j}, b_{2j}, \ldots, b_{nj})$ holds when $\boldsymbol{A}$ is a binary matrix, then by replacing $\sum_i \boldsymbol{z}_{ij}$ with $\boldsymbol{Z}_j \boldsymbol{a}_{\cdot j}$, a general solution of matrix B becomes to

$$\boldsymbol{B}^* = [\boldsymbol{Z}_1 \boldsymbol{a}_{\cdot 1} \quad \boldsymbol{Z}_2 \boldsymbol{a}_{\cdot 2} \quad \ldots \quad \boldsymbol{Z}_m \boldsymbol{a}_{\cdot m}]_{n \times m}, \tag{A5}$$

where $\boldsymbol{Z}_j = [\boldsymbol{z}_{1j} \quad \boldsymbol{z}_{2j} \quad \cdots \quad \boldsymbol{z}_{nj}]_{n \times n}$, and $\boldsymbol{a}_{\cdot j}$ is the jth column of matrix $\boldsymbol{A}$.

In solving this optimization problem, one of the main challenges is to handle the large data matrix $\boldsymbol{A}$. To be computationally efficient, we propose to use a *divide-and-conquer* algorithm. Specifically, we split the large matrix $\boldsymbol{A}$ row-wise into several smaller submatrices. For each submatrix, the optimal solution can be obtained efficiently. Then the solution to the original problem (matrix $\boldsymbol{A}$) is the combination of the solutions of the subproblems. A simple strategy for splitting matrix $\boldsymbol{A}$ could be random splits with equal sizes. Our simulation study demonstrates that the divide-and-conquer method approximates the original problem very well while significantly reduces the computational cost.

To formulate the "*l*-diversity" constraint (6) as a set of linear constraints, we introduce an additional binary decision variable $\boldsymbol{u}$ that is of length np. Then constraint (6) can be represented as

$$C_4 \boldsymbol{z} - \mathbf{I}_{np} \boldsymbol{u} + \mathbf{0} \boldsymbol{v} \geq \mathbf{0}; \tag{A6}$$

$$\mathbf{0} \boldsymbol{z} + C_5 \boldsymbol{u} - l \mathbf{I}_n \boldsymbol{v} \geq \mathbf{0}; \tag{A7}$$

$$\mathbf{0} \boldsymbol{z} + C_5 \boldsymbol{u} - q \mathbf{I}_n \boldsymbol{v} \leq \mathbf{0}; \tag{A8}$$

where C_4 is an np by ns matrix such that $C_4 \boldsymbol{z} = (\boldsymbol{b}_1 . \boldsymbol{S}_{dum}, \ldots, \boldsymbol{b}_n . \boldsymbol{S}_{dum})$, and $C_5 = \mathbf{I}_n \otimes \mathbf{1}_p$.

B. Simulation to Demonstrate Performance and Computational Efficiency of Heuristic

We conduct several simulation studies to demonstrate that (1) our heuristic gains significant computational efficiency at a modest cost of optimality in terms of loss of information, (2) the scale of subproblems can be reasonably small for various sizes of panel data, and (3) different levels of distortion can be achieved using our weighted protection scheme.

B.1. Investigating performance and computational cost

The proposed heuristic can efficiently solve the optimization problem by using the divide-and-conquer method. In this simulation, we demonstrate that computing time can be significantly reduced by splitting matrix $\boldsymbol{A}$ into multiple submatrices and solving the optimization problem for these submatrices. We further show that the substantial gain in computational efficiency comes at a modest loss of information measured by the same metrics as in Table 3.

We simulate purchases of four brands by 300 panelists on multiple transactions, where the purchase quantities of the four brands are assumed to be the QID. Specifically, for each panelist, the number of records (transactions) is a random integer in [3, 15] with equal probability. For each record, the choices of the four brands follow a multinomial distribution with probabilities (0.1, 0.2, 0.3, 0.4). To account for multi-brand purchases, we set the size of the multinomial distribution to be a random integer in{1, 2, 3} with equal probability; that is, a panelist can purchase at most 3 different brands in a single transaction. The purchase quantity is generated from a Gamma distribution (rounded) with shape=1 and scale=2.

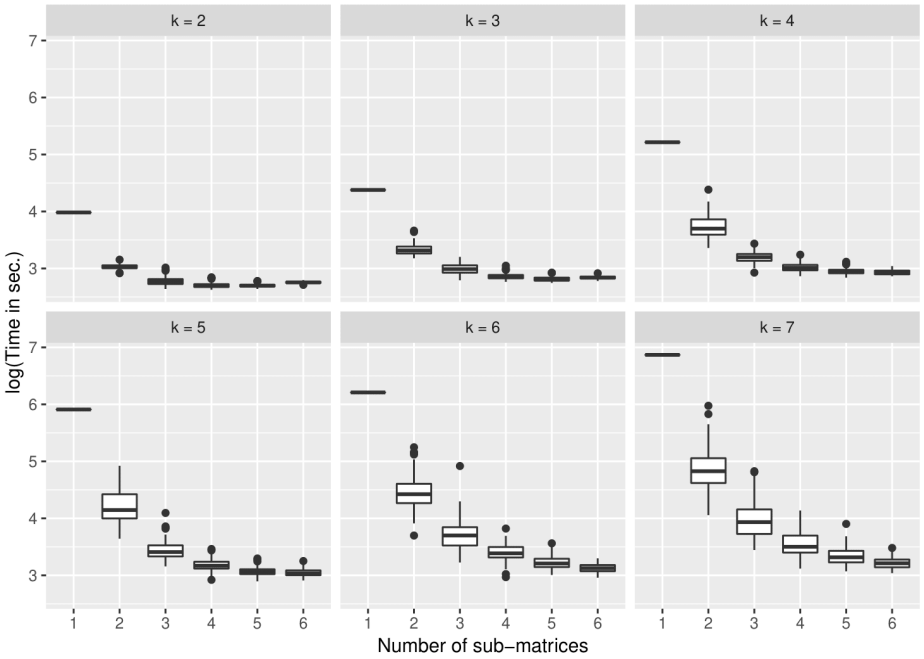

Figure A1. Computational time (in log seconds) for various privacy levels (k) and number of sub-matrices (q) split from the original matrix $\boldsymbol{A}$. The boxplots are for 100 replicates when q=2,…,6.

Figure A1 depicts the CPU time in natural log scale for k-anonymization across different k and q, which is the number of submatrices. For q>1, we repeat the k-anonymization 100 times due to randomness involved in the heuristic. The substantial reduction in CPU time as q increases can be clearly seen. For example, for k=5, solving the full optimization problem (q=1) would take about $e^6(\approx 400)$ seconds, while splitting into 6 sub-problems (q=6) and solving all of them (not in parallel computing) only takes $e^3(\approx 20)$ seconds. The computing time can be further reduced if parallel computing is applied. Figure A1 also shows that the computing time increases with the privacy level k. Similar results are observed based on a different random dataset.

The next results shown in Table A1 demonstrate that the gain of computational efficiency comes at a modest cost of optimality in terms of information loss. In particular, we compare the information loss based on solving the full optimization problem (q=1) and that of solving multiple (q=2,…,6) subproblems (divide-and-conqure). We report the same metrics as in Table 3 for 4-anonymization on four independently simulated samples based on the same setting as before. These include two measures for overall data deviation from the original true data (in panel A): the percentage of changed cells (PCC) and the mean squared distance (MSD),

and the mean absolute percentage deviation (MAPD) of three commonly reported marketing brand metrics (in panel B): market share (MS), share of category requirements (SCR) and brand switching probability (BSP). For q>1, we obtain the average and standard deviation across 100 replicates due to random split. We see an incremental loss of information as q increases for all measures except for BSP in a few cases. This shows the tradeoff between computational efficiency and the approximation error, which is expected in general. However, such incremental loss of information is modest compared to the reduced computing time as shown in Figure A1. Practitioners may carefully balance this tradeoff using context-specific metrics.

Table A1: Information loss due to 4-anonymization. Compared are the same metrics as in Table 3 for solving the full optimization problem (q=1) and solving multiple q (=2,…,6) sub-problems. Panel A reports overall information loss including percentage of changed cells (PCC) and mean squared distance (MSD) between protected and true data. Panel B reports mean absolute percentage deviation (MAPD) of market share (MS), share of category requirements (SCR) and brand switching probability (BSP). Reported are average and standard deviation (in parentheses) over 100 replicates for q=2…6.

		q=1	q=2	q=3	q=4	q=5	q=6
Panel A. Overall information loss in transaction-level data							
Data #1	PCC	2.85 (-)	3.27 (0.07)	3.56 (0.08)	3.74 (0.11)	3.91 (0.1)	4.01 (0.11)
	MSD	0.27 (-)	0.36 (0.02)	0.42 (0.02)	0.48 (0.03)	0.52 (0.03)	0.55 (0.03)
Data #2	PCC	3.02 (-)	3.45 (0.08)	3.71 (0.08)	3.91 (0.09)	4.09 (0.10)	4.18 (0.10)
	MSD	0.28 (-)	0.33 (0.01)	0.37 (0.01)	0.41 (0.01)	0.44 (0.01)	0.47 (0.02)
Data #3	PCC	3.14 (-)	3.69 (0.07)	4.03 (0.09)	4.24 (0.10)	4.41 (0.10)	4.57 (0.11)
	MSD	0.29 (-)	0.38 (0.01)	0.44 (0.02)	0.48 (0.02)	0.53 (0.02)	0.56 (0.02)
Data #4	PCC	2.85 (-)	3.27 (0.07)	3.56 (0.08)	3.74 (0.11)	3.91 (0.10)	4.01 (0.11)
	MSD	0.27 (-)	0.36 (0.02)	0.42 (0.02)	0.48 (0.03)	0.52 (0.03)	0.55 (0.03)
Panel B. MAPD (%) in brand metrics							
Data #1	MS	0.70 (-)	1.11 (0.38)	1.22 (0.41)	1.49 (0.52)	1.66 (0.59)	1.86 (0.58)
	SCR	0.31 (-)	0.55 (0.21)	0.66 (0.28)	0.83 (0.30)	0.94 (0.43)	1.09 (0.41)
	BSP	0.72 (-)	0.97 (0.31)	1.07 (0.29)	1.24 (0.31)	1.33 (0.38)	1.27 (0.36)
Data #2	MS	0.28 (-)	1.11 (0.39)	1.28 (0.49)	1.33 (0.42)	1.34 (0.50)	1.46 (0.53)
	SCR	0.45 (-)	0.60 (0.23)	0.64 (0.26)	0.74 (0.29)	0.74 (0.25)	0.75 (0.28)
	BSP	0.63 (-)	0.72 (0.23)	0.82 (0.26)	0.81 (0.23)	0.78 (0.27)	0.83 (0.29)
Data #3	MS	0.65 (-)	0.84 (0.32)	0.95 (0.34)	1.09 (0.41)	1.11 (0.43)	1.22 (0.44)
	SCR	0.42 (-)	0.55 (0.23)	0.63 (0.27)	0.62 (0.25)	0.67 (0.28)	0.78 (0.36)
	BSP	0.61 (-)	0.72 (0.24)	0.89 (0.29)	0.85 (0.32)	0.98 (0.34)	1.06 (0.31)
Data #4	MS	0.47 (-)	0.57 (0.27)	0.79 (0.34)	0.91 (0.36)	1.04 (0.44)	1.13 (0.49)
	SCR	0.32 (-)	0.53 (0.22)	0.63 (0.29)	0.62 (0.26)	0.74 (0.32)	0.78 (0.33)
	BSP	1.27 (-)	1.09 (0.24)	1.09 (0.22)	1.07 (0.28)	1.02 (0.29)	1.07 (0.33)

B.2. Scale of subproblems for different sizes of panel data

In this subsection, we assess the dimension of matrix $\boldsymbol{A}$ and its submatrices under different sizes of the original panel data. First, we generate panel data with different sizes (by varying the number of panelists

and number of brands), and transform it to matrix $\boldsymbol{A}$. We investigate the size of matrix $\boldsymbol{A}$ and its submatrices by splitting row-wise. Recall that the number of rows of matrix $\boldsymbol{A}$ is the number of unique QID values in the data, and number of columns of $\boldsymbol{A}$ is the number of panelists. Thus, the dimension of matrix $\boldsymbol{A}$ and the total number of nonzero entries determines the scale of the original optimization problem, which can be large and infeasible for a large dataset. The divide-and-conquer method in our heuristic is the key to improving the computational efficiency. The simulation results in B.1 have shown the tradeoff between the number of submatrices and the optimality. Therefore, the goal of this simulation is to investigate the dimension of submatrices across different sizes of the original data.

We use a similar data generating procedure as described in B.1 except for the following changes: (1) we vary the number of panelist in (100, 300, 500, 1000, 5000, 10000) and the number of brands (QID attributes) in (3, 4, 5, 6, 7, 8, 9) correspondingly; and (2) the brand shares are set to be equal, so the choices in each record (transaction) are simulated from a multinomial distribution with equal probabilities.

Table A2 shows the number of observations in the simulated data (# obs.), number of rows of matrix $\boldsymbol{A}$ (row(A)), and average number of columns of the q submatrices split from matrix $\boldsymbol{A}$. The reported values are averages over 100 replicates for different numbers of panelists (# ID). We can see that the average number of nonzero columns of submatrices decreases with increasing q. For data with up to 1000 panelists, the number of submatrices q does not increase beyond 10. The number of rows of submatrices is simply nrow(A)/q. Therefore, for sufficiently large M, we can always make the dimension of submatrices small enough to achieve certain computational efficiency. For instance, in Table A2 # ID = 300 corresponds to the panel data generated in previous simulation study (B.1). We can see that the average dimension of matrix $\boldsymbol{A}$ is 437 rows and 300 columns, while its submatrices when q=6 have 73 rows (437/6) and 213 columns on average, and the computational cost for this size of data is small enough as shown in Figure A1. For moderate to large size of panel data, e.g., 10,000 panelists and 9 different brand purchases, the average number of records in the panel data is 81,257 (# obs.) and the average number of unique QIDs (nrow(A)) is 7031. When q=80, the average dimension of a single submatrix is about 88 rows (7031/80) and 939 columns, which is of a moderate size to handle unlike the original 7031*10000 dimensional matrix $\boldsymbol{A}$. In practice, however, one may

carefully choose the number of splits q to balance the tradeoff between computational efficiency and optimality. In the next simulation, we demonstrate that a submatrix with this moderate size can be efficiently solved using a personal computer.[1]

Table A2: Average sizes of panel data, matrix $\boldsymbol{A}$ and submatrices for data with different numbers of panelists.

# ID	# obs.	row(A)	Avg. # nonzero columns of submatrices for different q								
			q=2	q=4	q=6	q=8	q=10	q=20	q=30	q=50	q=80
100	789.8	172.6	97.0	83.4	70.3	59.7	51.8	-	-	-	-
300	2392.5	436.7	291.9	251.4	212.6	181.6	158.1	-	-	-	-
500	4021.2	747.5	487.0	422.4	357.4	306.5	266.3	-	-	-	-
1000	8097.8	1369.8	974.6	847.3	719.6	616.8	538.1	-	-	-	-
5000	40506.2	4038.8	4874.0	4239.1	3602.8	3094.6	2691.1	1614.8	1146.1	724.9	466.7
10000	81257.2	7031.1	9752.9	8499.5	7226.3	6208.7	5410.1	3247.4	2308.2	1459.6	939.4

To demonstrate, we directly simulate a submatrix of dimension 100*500 and 100*1000,[2] and apply our k-MM method at k=7 to each setting, respectively. We find that the average CPU time for the two different sizes of submatrix are 7.14 (0.81) and 12.2 (0.58) seconds (shown in parentheses are standard deviations over 100 replicats). It indicates that for a panel data with 10000 panelists as simulated previously, our method takes approximately 16 minutes for 7-anonymization when q=80, without parallelizing.

B.3. Weighted cost of distortion

Following the same data generating procedure described in B.1, we simulate 100 samples and apply the weighted protection scheme described in Section 4.2 to show that different levels of distortion can be achieved for different attributes. Figure A2 shows the boxplot of the MAPD of sales of all four brands under weighted (right) and unweighted (left) protection schemes for different privacy levels k. The boxplot displays results for 100 random samples. This result is consistent with Figure 9, which shows the MAPD of sales for real household panel data. Again, the results demonstrate that the proposed data protection method allows the flexibility that user can specify different levels of distortion for different attributes.

[1] This simulation is performed on a Dell XPS 8930 with Intel Core i9-9900K CPU and 32 GB memory.
[2] The entries of each column (length=100) are generated from a multinomial distribution, where the vector of probabilities is c(1,…,1, 2,…,2, 3,…,3, 4,…,4, 5,…,5), where c is a scaling constant such that the sum of this vector is 1, and the size is 1 (with probability 0.67) or 2 (with probability 0.33)

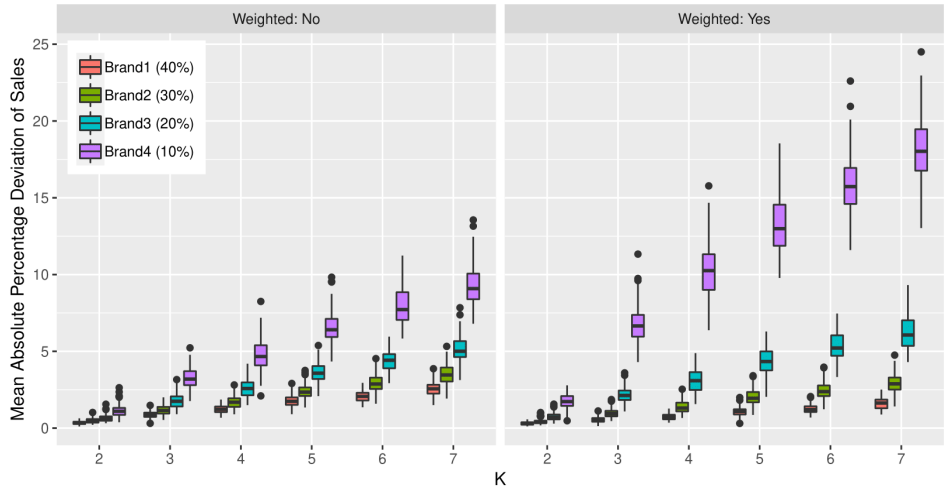

Figure A2. Mean absolute percentage deviation of sales for four different brands under weighted and unweighted protection scheme over 100 samples. The weights are the true market shares (shown in parentheses in the legend).

C. Brand Choice Model and Data Processing

Brand Choice Model

In this paper, we focus on brand choice modeling (McFadden, 1974; Guadagni and Little, 1983; Gupta et al., 1996; Allenby and Rossi, 1998) as an important marketing application to evaluate data utility after protection. The Hierarchical Bayesian random coefficient model (Allenby and Rossi, 1998) is a popular approach for estimating the brand choice model with customer heterogeneity.

Let $\boldsymbol{y}_{it}^{(j)}$ be an observed $J \times 1$ indicator vector where jth element is 1 and others are 0, indicating that customer i chooses brand j over other brands at time t. Under the standard random utility framework, the brand choice for customer i at time t, denoted by the random vector Y_{it}, can be modeled as a function of latent utility of all alternatives $(U_{it}^{(1)}, \ldots, U_{it}^{(J)})$. Specifically,

$$\mathbb{P}\left(Y_{it} = \boldsymbol{y}_{it}^{(j)}\right) = G\left(U_{it}^{(j)}\right); \tag{A6}$$

$$U_{it}^{(j)} = \alpha_{ij} + \boldsymbol{\beta}_i^T \boldsymbol{x}_{ijt} + \epsilon_{ijt}, \quad \boldsymbol{\epsilon}_{i,t} \overset{iid}{\sim} N(\mathbf{0}, \Lambda), \tag{A7}$$

where the parameter α_{ij} in (A7) is customer i's intrinsic brand preference for brand j, and $\boldsymbol{\beta}_i^T \boldsymbol{x}_{ijt}$ is the systematic component, where the coefficient vector $\boldsymbol{\beta}_i \in \mathbb{R}^K$ reflects the effect of marketing mix variables X (e.g. price, promotions, display and feature) for customer i. Commonly used link function $G^{-1}(\cdot)$ is

multinomial logit or probit model. Under the hierarchical Bayesian modeling framework, the parameter vector $(\alpha_{i1}, \ldots, \alpha_{iJ}, \boldsymbol{\beta}_i) \in \mathbb{R}^{(J+K)}$ is the random coefficient vector that is assumed to independently and identically follow a multivariate normal distribution. Denoting $\boldsymbol{\theta}_i = (\alpha_{i1}, \ldots, \alpha_{iJ}, \boldsymbol{\beta}_i)$, equation (A7) can be expressed as the following hierarchical Bayesian random effect model:

$$U_{it(j)} | \boldsymbol{\theta}_i = \boldsymbol{\theta}_i^T \boldsymbol{x}_{ijt}^* + \epsilon_{ijt}, \quad \boldsymbol{\epsilon}_{i,t} \overset{iid}{\sim} N(\mathbf{0}, \Lambda); \tag{A8}$$

$$\boldsymbol{\theta}_i \sim \mathrm{N}(\overline{\boldsymbol{\theta}}, V_{\boldsymbol{\theta}}), \tag{A9}$$

where $\boldsymbol{x}_{ijt}^*$ is $(J + K) \times 1$ covariate vector with first J elements being 0 except for jth being 1. Following Allenby and Rossi (1998), we assume conjugate priors for the mean vector $\overline{\boldsymbol{\theta}}$, covariance matrix $V_{\boldsymbol{\theta}}$, and Λ. That is $\overline{\boldsymbol{\theta}} \sim \mathrm{N}(\overline{\overline{\boldsymbol{\theta}}}, aV_{\boldsymbol{\theta}})$, $V_{\boldsymbol{\theta}} \sim \mathrm{IW}(u_0 I, v_0)$, where $v_0 > K + J$, and $\Lambda = \mathrm{diag}(1, \lambda_2, \ldots, \lambda_{J-1})$ with $\sqrt{\lambda_j} \sim \mathrm{IG}(\nu, s_j)$, where $\nu = 3$ and $s_j = 1$. Note that λ_1 is set to 1 for identifiability in the multinomial model.

This model can be efficiently estimated through Markov Chain Monte Carlo methods with Gibbs sampling. We use the R package "bayesm" in our empirical analysis.

Data Preparation

Each marketing mix variable in the store data file is first aggregated over UPCs for each of the 10 brands, to obtain a week-by-brand table. This table is merged into the panel purchasing data by week. The process of aggregation is to compute volume-weighted average prices across UPCs, and the mode across UPCs of each binary promotion indicator. Next, only for the brand purchased on each shopping trip, the marketing mix variable value in the panel data overwrites the value obtained from the store data. Estimation of the brand choice model with unobserved heterogeneity requires that each household should have multiple purchase records. Therefore, we include in our sample only households who have made 5 or more purchases in the data. We exclude 14.9% of trips on which multi-brand purchases occur. As a result, the sample for estimating the brand choice model has 3,995 shopping trips made by 407 households. It is important to clarify that application of these filters may result in slightly different sample sizes for each protected data. This is because the protection is applied to the original data and it alters the number of units purchased, in which case the alteration from nonzero to zero (or vice versa) would affect the filtering results. However, Table A7

in the appendix D shows that the differences are negligible. It is important to note that the data protection procedure is applied to the full sample, while brand choice modeling is based on this filtered subsample.

D. Additional Results

Table A3: Unicity for increasing numbers of top brands for 15 product categories. The QID is defined as the set of attributes: week and number of units purchased of each brand.

Product category	Unicity for top J brands							
	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
Carb beverage	0.348	0.402	0.461	0.490	0.539	0.587	0.621	0.640
Coffee	0.184	0.264	0.293	0.315	0.347	0.362	0.371	0.387
Cold cereal	0.187	0.204	0.219	0.252	0.311	0.344	0.369	0.406
Frozen dinner	0.452	0.374	0.398	0.398	0.416	0.409	0.424	0.458
Frozen pizza	0.271	0.301	0.327	0.346	0.350	0.369	0.382	0.390
Hotdog	0.113	0.142	0.161	0.198	0.211	0.227	0.247	0.262
Laundry detergent	0.161	0.168	0.181	0.200	0.205	0.224	0.238	0.247
Mayo	0.120	0.123	0.126	0.132	0.136	0.137	0.140	0.144
Milk	0.097	0.127	0.152	0.162	0.185	0.195	0.204	0.208
Mustard/Ketchup	0.095	0.117	0.151	0.169	0.203	0.214	0.233	0.247
Peanut butter	0.111	0.137	0.154	0.167	0.170	0.179	0.183	0.199
Salty snack	0.140	0.183	0.235	0.276	0.320	0.355	0.381	0.414
Spaghetti sauce	0.147	0.183	0.212	0.221	0.244	0.259	0.269	0.287
Toilet tissue	0.078	0.091	0.103	0.117	0.137	0.148	0.159	0.167
Yogurt	0.309	0.361	0.404	0.431	0.462	0.484	0.502	0.509

Table A4: Sno-unicity for increasing numbers of top brands for 15 product categories. The QID is defined as the set of attributes: week and number of units purchased of each brand.

Product category	Sno-unicity for top J brands							
	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
Carb beverage	0.631	0.725	0.805	0.850	0.904	0.928	0.939	0.942
Coffee	0.278	0.417	0.443	0.470	0.508	0.535	0.558	0.586
Cold cereal	0.331	0.378	0.375	0.431	0.530	0.586	0.629	0.664
Frozen dinner	0.596	0.509	0.563	0.557	0.604	0.590	0.609	0.637
Frozen pizza	0.416	0.470	0.489	0.514	0.514	0.533	0.563	0.580
Hotdog	0.159	0.208	0.243	0.278	0.308	0.326	0.348	0.361
Laundry detergent	0.217	0.242	0.261	0.285	0.291	0.311	0.325	0.339
Mayo	0.153	0.157	0.159	0.163	0.167	0.168	0.171	0.175
Milk	0.187	0.224	0.280	0.295	0.326	0.334	0.337	0.345
Mustard/Ketchup	0.104	0.136	0.177	0.207	0.255	0.275	0.302	0.324
Peanut butter	0.153	0.204	0.238	0.265	0.268	0.282	0.286	0.297
Salty snack	0.248	0.337	0.425	0.494	0.585	0.681	0.724	0.775
Spaghetti sauce	0.208	0.265	0.312	0.335	0.369	0.380	0.392	0.411
Toilet tissue	0.131	0.200	0.208	0.218	0.256	0.275	0.301	0.305
Yogurt	0.606	0.721	0.777	0.810	0.835	0.848	0.851	0.856

Table A5: Cumulative market shares of top J brands for 15 product categories.

Product category	Cumulative market share of top J brands									
	J=1	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10

Carb beverage	0.157	0.251	0.311	0.370	0.418	0.466	0.514	0.555	0.583	0.605
Coffee	0.287	0.428	0.526	0.606	0.683	0.720	0.748	0.774	0.799	0.820
Cold cereal	0.060	0.118	0.172	0.208	0.243	0.275	0.306	0.335	0.365	0.392
Frozen dinner	0.121	0.239	0.345	0.435	0.478	0.516	0.553	0.587	0.621	0.648
Frozen pizza	0.196	0.385	0.506	0.627	0.701	0.757	0.801	0.836	0.868	0.896
Hotdog	0.378	0.515	0.645	0.755	0.857	0.879	0.897	0.911	0.925	0.938
Laundry detergent	0.314	0.626	0.754	0.827	0.884	0.940	0.952	0.964	0.971	0.975
Mayo	0.760	0.943	0.980	0.985	0.988	0.991	0.993	0.995	0.997	0.998
Milk	0.382	0.744	0.831	0.908	0.937	0.956	0.961	0.967	0.972	0.976
Mustard/Ketchup	0.410	0.783	0.830	0.863	0.893	0.924	0.943	0.958	0.970	0.979
Peanut butter	0.546	0.703	0.857	0.902	0.938	0.949	0.958	0.966	0.974	0.981
Salty snack	0.130	0.230	0.306	0.371	0.436	0.488	0.539	0.581	0.623	0.661
Spaghetti sauce	0.383	0.505	0.626	0.725	0.818	0.905	0.932	0.943	0.953	0.961
Toilet tissue	0.302	0.477	0.623	0.734	0.816	0.869	0.904	0.934	0.963	0.983
Yogurt	0.281	0.514	0.623	0.731	0.769	0.806	0.833	0.855	0.870	0.883

Table A6: Coefficient estimates of a logistic regression that predicts individual privacy risk (IR) coded as high (1) versus low (0) based on median split using household-specific purchase behavior characteristics.

	Estimate	Standard error	z value	p-value
Intercept	-1.366	0.187	-7.304	0.000
Avg. # of units bought per trip	0.234	0.070	3.363	0.001
# Brands ever bought	0.114	0.049	2.330	0.020
Total dollar spend across trips	0.010	0.002	4.461	0.000

Table A7: Average sample size and number of households over 100 replicates of protected data (filtered for brand choice model). In parentheses are standard deviations based on the 100 replicates.

	Unprotected	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
Average sample size	3995.0	3993.5 *(3.9)*	4012.1 *(4.7)*	4014.8 *(6.1)*	4031.4 *(7.6)*	4065.7 *(11.1)*	4084.1 *(11.3)*
Average number of panelists	407.0	406.8 *(0.5)*	408.1 *(0.7)*	408.4 *(0.8)*	408.3 *(1.1)*	409.6 *(1.4)*	410.2 *(1.3)*

Table A8: Brand-specific constant estimates.

	Unprotected	*k*-anonymized Data					
Brand-specific constants:	Data	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
Cheetos	-0.778	-0.790	-0.753	-0.723	-0.800	-0.809	-0.852
Doritos	-2.033	-2.177	-2.128	-2.111	-2.156	-2.139	-2.141
Lays	-0.569	-0.866	-0.923	-0.877	-0.902	-0.957	-1.055
Lays Natural	-2.138	-2.252	-2.253	-2.248	-2.271	-2.279	-2.328
Old Dutch	-0.095	-0.064	-0.023	-0.036	-0.066	-0.061	-0.124
Old Dutch Ripples	-3.424	-3.449	-3.358	-3.336	-3.421	-3.391	-3.424
Sunchips	-3.320	-3.396	-3.344	-3.331	-3.366	-3.311	-3.344
Tostitos Natural	-3.985	-4.145	-4.108	-4.097	-4.035	-4.048	-4.047
Wavy Lays	-2.275	-2.298	-2.263	-2.300	-2.326	-2.329	-2.388

Table A9: Standard deviations of posterior distributions for coefficient estimates.

	Unprotected	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
Cheetos	0.957	1.013	1.023	0.980	0.942	0.959	0.942
Doritos	1.204	1.223	1.196	1.178	1.150	1.133	1.108

Lays	1.181	1.282	1.289	1.248	1.234	1.231	1.184
Lays Natural	1.358	1.426	1.409	1.402	1.366	1.360	1.325
Old Dutch	0.953	0.893	0.905	0.902	0.900	0.884	0.874
Old Dutch Ripples	0.957	1.070	1.081	1.054	1.059	1.024	1.029
Sunchips	1.410	1.583	1.605	1.587	1.561	1.541	1.512
Tostitos Natural	1.500	1.524	1.525	1.521	1.491	1.496	1.469
Wavy Lays	1.223	1.112	1.081	1.054	1.073	1.045	1.023
LogPrice	1.500	1.603	1.624	1.612	1.574	1.552	1.529
Promotion	0.636	0.700	0.712	0.718	0.706	0.706	0.708
Display	0.826	0.822	0.829	0.834	0.838	0.842	0.823
Feature	0.701	0.770	0.801	0.749	0.744	0.743	0.725

Table A10: Sample size and number of households due to *k*-anonymization with record deletion

	Unprotected	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
Full sample without filtering							
# obs.	6117	5143	4642	4209	3779	3430	3130
# Panelists	1009	981	965	945	908	891	862
Filtered sample for brand choice modeling							
# obs.	3995	3630	3231	2785	2407	2082	1759
# Panelists	407	391	372	333	296	266	229

E. Implementation of Benchmarking Approaches

Implementation of clustering-based aggregation:

Input: Household panel data at transaction-level

Result: *k*-anonymized household panel data at transaction-level

For week t, do

1. Find m_t approximately equal-sized clusters based on QID variables (without week). To achieve k-anonymity, $m_t \leq \frac{n_t}{k}$, where n_t is the number of unique panelists in week t.
2. For each cluster, aggregate the QID variables with summation and market-mix variables with average, while keep the data at transaction-level, e.g., all n_t have the same transactions.

Definition of ϵ-differential privacy (Dwork 2006): A randomized function K gives ϵ-differential privacy if for all data sets D1 and D2 differing on at most one element and all measurable subsets $S \subseteq Range(K)$,

$\Pr[K(D_1) \in S] \leq exp(\varepsilon) \times \Pr[\mathrm{K}(D_2) \in \mathrm{S}]$.

Implementation of ϵ-differential privacy-based random noise addition:

Input: Household panel data at consumer-week-brand level

Result: Protected household panel data at consumer-week-brand level

1. Let $(n_1, n_2, \dots, n_J)$ be the total counts for the J brands. Denote $n \coloneqq \sum_1^J n_j$.

---

2. Set the hyperparameter of Dirichlet prior $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_J)$, and let $\alpha_1 = \alpha_2 = \cdots = \frac{n}{\exp(\epsilon)-1}$ where $\epsilon \geq 0$ is the level of protection.
3. For each brand j, sample one vector of posterior probabilities (consisting of J elements) from the posterior Dirichlet distribution with parameter vector $(n_1 + \alpha_1, \ldots, n_J + \alpha_J)$, which results from the combination of the multinomial likelihood and Dirichlet prior.
4. For each observation i, draw a synthetic brand choice from a multinomial distribution with the posterior probabilities from 3.

F. Descriptive Statistics of Physician Prescribing Data

Table A11: Summary statistics of physician prescription data at physician-month level.

	Prescriptions				Detailing
	% share	Total Number	Mean	Variance	Mean frequency
Lipitor	34.33	5148	0.575	1.148	0.833
Zocor	21.04	3155	0.352	0.714	1.097
Crestor	17.67	2650	0.296	0.783	0.615
Non Drug Treatment	26.96	4042	0.451	1.117	-

Table A12: Mean values of parameter estimates from Hierarchical Bayesian random effects logit model for physician prescribing data. In parentheses are posterior standard deviations.

	Unprotected	*k*-Anonymized Data					
		$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
Lipitor	0.381 (0.308)	0.403 (0.303)	0.408 (0.303)	0.407 (0.303)	0.414 (0.302)	0.419 (0.302)	0.421 (0.302)
Zocor	-0.238 (0.049)	-0.222 (0.035)	-0.226 (0.033)	-0.230 (0.032)	-0.225 (0.033)	-0.230 (0.035)	-0.231 (0.034)
Crestor	-0.435 (0.056)	-0.433 (0.038)	-0.444 (0.037)	-0.458 (0.037)	-0.460 (0.038)	-0.465 (0.037)	-0.470 (0.037)
Detailing (market mix)	0.308 (0.015)	0.303 (0.013)	0.303 (0.013)	0.303 (0.013)	0.302 (0.013)	0.302 (0.013)	0.302 (0.013)
MAPD (%)	-	5.603	6.008	6.360	7.497	7.653	7.924

References

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1-2):57–78.

Dwork, Cynthia (2006), "Differential Privacy," 33rd International Colloquium on *Automata, Languages and Programming*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, eds., Springer Berlin Heidelberg, 1-12.

Guadagni, P. M., & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3), 203-238.

Gupta, S., Chintagunta, P., Kaul, A., & Wittink, D. R. (1996). Do household scanner data provide representative inferences from brand choices: A comparison with store data. *Journal of Marketing Research*, 33(4), 383-398.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, 105-142.