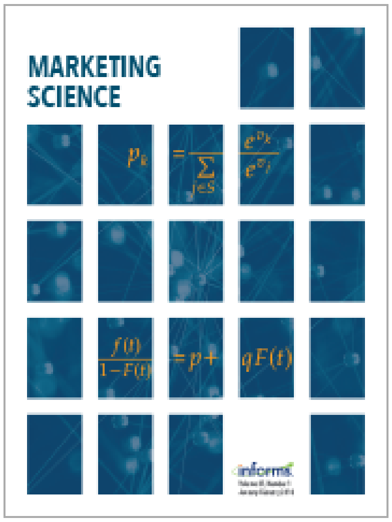


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Scalable Data Fusion with Selection Correction: An Application to Customer Base Analysis

Daniel Minh McCarthy , Elliot Shin Oblander

informs
http://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 40, No. 3, May-June 2021, pp. 459–480
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Scalable Data Fusion with Selection Correction: An Application to Customer Base Analysis

Daniel Minh McCarthy,[a] Elliot Shin Oblander[b]

[a] Department of Marketing, Emory University, Atlanta, Georgia 30322; [b] Marketing Division, Columbia University, New York, New York 10027

Contact: daniel.mccarthy@emory.edu, https://orcid.org/0000-0002-2366-2598 (DMM); EOblander23@gsb.columbia.edu, https://orcid.org/0000-0001-5310-8953 (ESO)

Received: July 23, 2019
Revised: March 17, 2020; May 17, 2020; May 28, 2020
Accepted: June 5, 2020
Published Online in Articles in Advance: February 2, 2021

https://doi.org/10.1287/mksc.2020.1259

Abstract. Increasingly, applied researchers study problems for which multiple sources of data are available. These sources may come with varying degrees of aggregation, and some of them may not be representative of the population of interest. Using multiple data sources could lead to richer insights. However, existing data fusion approaches do not correct for selection bias in data sources that may not be representative and either do not scale to large populations or are statistically inefficient. We propose an aggregate-disaggregate data fusion method that corrects for selection bias and is both computationally scalable and statistically efficient. We apply the method to estimate a model of customer acquisition and churn at subscription-based firms. We bring the model to life using a large credit card panel and public data from Spotify, the music streaming service. This application and supporting simulations show that incorporating the granular data through our data fusion method enhances identification and offers richer insights than extant approaches. We find, for example, that previously churned customers remain with Spotify longer than newly adopted subscribers do, implying a more sanguine view of Spotify's future retention profile than previous approaches that do not use multiple data sources.

History: Avi Goldfarb served as the senior editor and Christophe Van den Bulte served as associate editor for this article.
Funding: The work is partially supported by Emory University's Dean's Research Grant and Adobe's Digital Marketing Research Award.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mksc.2020.1259.

Keywords: data fusion • selection correction • customer relationship management • marketing-finance interface

1. Introduction

The diversity of data sources has increased significantly in recent years. At last count, the website ProgrammableWeb provides searchable access to approximately 23,000 application programming interfaces (APIs), facilitating the use of a broad array of granular data sets. Third-party data companies sell access to an equally diverse collection of data sets. Although household purchase and viewership data from companies such as Kantar, Comscore, Nielsen, and IRI are relatively popular in the academic marketing literature, there has been a groundswell of other more modern, less well-known data sets, including credit card panel data (Second Measure), geolocation data (PlaceIQ, Mogean), clickstream data (jumpshot), email receipt data (Rakuten Intelligence), and more. These data sources supplement traditional census-level data sources, including aggregate data collected by and/or filed with governmental institutions and industry groups (e.g., the Securities and Exchange Commission (SEC), Census Bureau, and National Retail Federation).

As a result, it is increasingly common that there is more than one data set available covering a particular population of interest. Modelers may naturally want to perform analysis using a *fusion* of data sources, but doing so creates new methodological challenges because of the difficulty of incorporating differing granularities of data and selection bias. We propose a methodology to solve these challenges, allowing the fusing together of (1) aggregated data about the population as a whole with (2) granular data for a possibly nonrepresentative subsample of that population, when the size of the population may be very large. When a representative sample of granular data is not available, training a model on aggregated population-level data and granular data from a possibly nonrepresentative sample could allow modelers to derive the benefits of both sources—representativeness and rich individual-level visibility, respectively—while ameliorating their limitations.

This underlying data structure—representative but limited aggregated data and detailed but possibly nonrepresentative granular data—is an increasingly common one in marketing, economics, and finance. As mentioned previously, a large and growing number of data sets are now available to marketing researchers, increasing the range of questions these researchers can answer, subject to the availability of suitable methods.

The aggregate-disaggregate data fusion problem has previously been studied by Feit et al. (2013), who build off of prior work on Bayesian estimation of individual-level models of consumer choice using only aggregated data (Chen and Yang 2007). Similar data fusion problems have arisen in a variety of fields outside of marketing, such as fusing aggregate product market shares with household-level survey data in economics (Berry et al. 2004) or fusing aggregate census demographic data with disaggregate trip-level data in transportation research (Dias et al. 2019). Despite the prevalence of this data structure across disciplines, extant literature proposing generally applicable methodologies for aggregate-disaggregate data fusion is limited; what methods have been proposed assume that there is no selection bias in the disaggregate data, do not scale to large populations, and/or require potentially arbitrary aggregation of granular data into summary statistics.

We propose a computationally scalable estimator for aggregate-disaggregate data fusion that is able to correct for selection bias in the disaggregate data. We achieve scalability by asymptotically approximating the likelihood of the observable data with a *proxy likelihood* function that is efficient to compute. The intuition behind our approximation is as follows. The exact likelihood of the aggregate data is generally not feasible to compute directly, making it difficult to incorporate aggregate data into a likelihood-based estimation procedure. However, for large target populations, many types of aggregate summary statistics—including sample averages and transformations of them—will converge to a normal distribution. Thus, we can approximate the likelihood of the aggregated data using a normal likelihood. This allows for fast approximation of the likelihood function, because the normal approximation requires only the first two moments of the aggregate statistics (analogous to common moment-based estimators), making it feasible to approximately maximize the joint likelihood of the aggregate and disaggregate data.[1] Within this computational procedure, we correct for selection bias by allowing the distribution of heterogeneous characteristics (e.g. individual-level parameters in a mixture model) to differ between the individuals underlying the disaggregate data and the overall population of interest.

In sum, our contribution is to propose a general methodology for estimating a model using representative but limited aggregate data and disaggregate but possibly nonrepresentative panel data at scale to obtain the best of both worlds: richer inferences and more accurate predictions than if we had used either data source on its own.

1.1. Scope and Limitations

Before discussing our specific customer base analysis use case, we first explicitly delineate the general applicability, benefits, and limitations of our method. In doing so, we illustrate not only the usefulness of the methodology beyond our specific empirical application but also the boundary conditions under which it may not be beneficial.

The specific aggregate-disaggregate data fusion problem to which our method is applicable is the estimation of an individual-level *micro*-statistical model for a population of interest (e.g., all households in the United States),[2] for which researchers have at their disposal a combination of (1) aggregate *macro* data that are representative of the entire population of interest and (2) disaggregate *micro* data that cover a possibly nonrepresentative subsample of that population. The aggregate data may include any summary statistic that is asymptotically normal, including but not limited to sample moments (under the central limit theorem), nonlinear transformations thereof (by the delta method), and sample quantiles. Our methodology allows for statistically and computationally efficient estimation in such settings, using both data sources jointly, while correcting for potential selection bias in the disaggregate data source.

Performing this type of data fusion can confer several advantages to researchers relative to estimating a model on only one of the two data sources:

- Compared with estimation using only limited aggregated data (e.g., customer base models estimated on aggregate customer base statistics reported to the SEC), incorporating disaggregate data provides two distinct benefits: first, for a given model specification, the added information can improve statistical precision, leading to more accurate inferences and predictions; second, additional visibility into individual-level behavior can enable researchers to estimate more complex models that would not be identifiable with the available aggregate data alone, leading to richer insights into the underlying individual-level processes (Berry et al. 2004).
- Compared with estimation using only nonrepresentative disaggregate data (e.g., consumer choice models estimated on scanner panel data), incorporating the representative aggregate data allows researchers to generalize from the disaggregate data sample to the population of interest as a whole, achieving external validity in their estimates.

Although our proposed method enables researchers to reap these benefits in many application areas, there are cases where our method may not be beneficial.

Our method introduces the ancillary problem of estimating a model of selection into the disaggregate data; this entails a tradeoff between the added information about the population conveyed through the disaggregate data and the added estimation variance introduced by the selection model. When there is severe selection bias, the size of the disaggregate data is small, and/or the aggregate data are already so rich that the model is well identified using the aggregate data alone, the incremental information gained through the disaggregate data may be small, so the costs could outweigh the benefits. Thus, our method is most likely to be beneficial in cases where models would only be weakly identified through aggregate data alone and where the disaggregate data are at least somewhat representative of the population of interest. Although our proposed method may improve identification relatively speaking, it does not guarantee strong identification in an absolute sense; indeed, in our empirical application in Section 6, model performance improves when incorporating disaggregate data, but some of the resulting standard errors are still fairly large.

Having provided an application-agnostic view of the strengths, weaknesses, and applicability of the proposed method, we narrow our focus in the next section to the specific setting to which we apply our method in this paper.

1.2. Application to Modeling Subscriber Acquisition and Retention

To frame the discussions of our model and methodology in Sections 2 and 3, we briefly describe and motivate our empirical application and the specific data structure that we encounter in it. We apply the proposed data fusion methodology to a customer base analysis problem: modeling customer acquisition and churn behavior at a subscription-based firm as if we were an external stakeholder. We analyze the quality and quantity of the customer base of the music streaming service Spotify, and how it has evolved over time.

This modeling exercise is arguably the most important step in the process of linking customer-level activity to the overall financial valuation of firms, commonly referred to as customer-based corporate valuation (CBCV). Gupta et al. (2004) provided the first proof of concept for how CBCV could be implemented for publicly traded firms. A large number of papers in marketing have built on this seminal work, studying the valuation implications of firm capital structure (Schulze et al. 2012), heterogeneous customer retention (McCarthy et al. 2017), business type (McCarthy and Fader 2018), and more. Other disciplines have written papers on this topic as well, including finance (Gourio and Rudanko 2014) and accounting (Bonacchi et al. 2015).

As in prior literature, we model customer acquisition and retention through a series of hazard models governing (1) the duration of time until customers are acquired and (2) how much time elapses after that before they churn. We assume that the analyst is an external stakeholder (e.g., an investor), and as such, only has access to external data sources (e.g., company data publicly disclosed through SEC filings and data from third-party providers) and not internal ones (e.g., internal customer relationship management system information). This *outside-in* perspective facilitates the valuation of market-based assets for investors, who ultimately determine the value of such assets through the financial markets. That said, similar data structures could arise in *inside-out* analyses, as we will discuss in Section 7.

There are two research gaps within extant CBCV literature that we address through this empirical application. The first gap is the range of the input data used in CBCV models. All the aforementioned papers only use aggregated customer data summaries disclosed by the companies themselves (e.g., the total number of customers acquired in a particular quarter). In addition to limiting the richness of the models we can specify, this limits analyses to firms that voluntarily disclose customer metrics on a regular basis, because public customer data disclosure is not mandatory (Bayer et al. 2017).

The second gap is the treatment of repeat acquisition and churn. Although customers who churn from a firm may be reacquired in future periods (and then churn again), none of the aforementioned papers separately model initial and repeat behaviors, because the resulting models would be difficult to identify from aggregated data alone. Repeat customer behavior is important for long-run firm outcomes, which are a primary driver of corporate valuation. As a company matures and the composition of its customer base shifts toward reacquired customers, its overall retention curve will shift from its initial retention curve to its repeat retention curve. This dynamic makes it important to know how the churn profile for newly acquired customers differs from that of reacquired customers. For instance, we show in our empirical example that Spotify's repeat retention curve is substantially higher than its initial retention curve, implying improving retention as Spotify matures. Despite the importance of capturing repeat behavior, all previous papers have been unable to separate out initial and repeat behaviors because of limited data.

As discussed in Section 1.1, these gaps are precisely those that are likely to be improved by data fusion: By supplementing limited aggregate data with rich disaggregate data, we can identify acquisition and churn models for more companies and can separate out initial and repeat behavior. Accordingly, we estimate our model using two data sources instead of one. As with extant literature, our first data source is a collection of aggregate summary statistics disclosed by Spotify itself through SEC filings and investor reports about its customer base. Our second data source is a credit card panel from alternative data firm Second Measure. Through this panel data, we can see monthly credit card spends at Spotify for each of approximately 3 million credit card panel members starting in January 2015. Although Second Measure's data set is large and granular, it is a nonrepresentative subsample of the overall population and covers only a part of Spotify's tenure.

Beyond Spotify, this data structure is very applicable to the CBCV use case. The aggregate summary statistics that Spotify disclosed are also disclosed by scores of other publicly traded companies, including those analyzed in prior literature. Furthermore, the credit card panel has company names associated with each purchase, making it a useful data source for all business-to-consumer companies. As such, the proposed methodology and data sources could be used for many other companies. We further discuss how this approach could be applied to other problems, both in CBCV and beyond, in Section 7.

The rest of the paper is organized as follows. We specify a model of customer acquisition and retention in Section 2. We describe our proposed methodology, with which we estimate the proposed model using aggregate and disaggregate data, in Section 3. We discuss identification of our model in Section 4 and then run a simulation study to understand the performance of the proposed methodology relative to extant approaches in Section 5. We provide an empirical analysis that applies the proposed estimation procedure to data from Spotify in Section 6 and then close with a discussion in Section 7. We include proofs of asymptotic properties, other derivations, and data in the online appendix.

2. Modeling Customer Acquisition and Retention

2.1. Individual-Level Model

To model customer dynamics in a subscription-based setting, we must specify the processes by which prospective customers (prospects) are acquired and the process by which they churn. Furthermore, a single individual can cancel a subscription and subsequently resubscribe, sometimes several times, and as such, it is important to model the process by which previously churned customers are reacquired. We specify these processes at the individual level in this section and then discuss how we account for unobserved heterogeneity in the next section. For notational simplicity, we denote the initial acquisition, initial churn, repeat acquisition, and repeat churn processes as IA, IC, RA, and RC, respectively.

We model time-to-acquisition and time-to-churn using a proportional hazards model with Weibull baseline hazard, a widely used duration model in customer base analysis and beyond (Schweidel et al. 2008b, McCarthy et al. 2017). In our empirical application, the customer payment cycle is monthly, so we discretize to the monthly level. Assuming that the value of a covariate is constant within each time period, the probability of not yet having been acquired m months after becoming a prospect (analogously, not yet having churned m months after becoming a customer) is

$$S(m|\lambda, c, \boldsymbol{\beta}, \mathbf{x}_{1:m}) = \exp(-\lambda B(m|c, \boldsymbol{\beta}, \mathbf{x}_{1:m})), \quad \text{for } m = 1, 2, \ldots \tag{1}$$

where

$$B(m|c, \boldsymbol{\beta}, \mathbf{x}_{1:m}) = \sum_{t=1}^{m} [t^c - (t-1)^c] \exp(\boldsymbol{\beta}' \mathbf{x}_t) \tag{2}$$

and $\mathbf{x}_{a:b}$ indicates the set of all variables $\mathbf{x}$ with indices in the range $a, a+1, \ldots, b$ (e.g., $\mathbf{x}_{1:3}$ represents $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$). Here, $\lambda > 0$ is a scale parameter, $c > 0$ is a shape parameter, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\mathbf{x}_t$ is a vector of covariates for month t.[3]

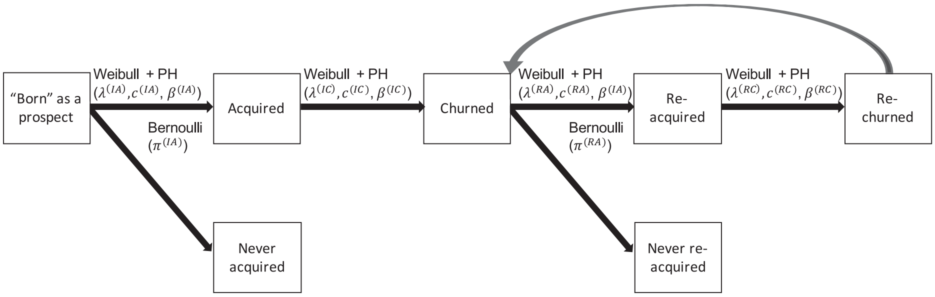
We allow each process to have distinct parameters: For example, there are four rate parameters $\lambda^{(IA)}, \lambda^{(IC)}, \lambda^{(RA)}$, and $\lambda^{(RC)}$ (likewise for $\boldsymbol{\beta}$ and c). In our empirical application, we use a vector of quarterly dummies as $\mathbf{x}_t$ for each of the four processes to account for seasonality. Following prior literature (Gupta et al. 2004, McCarthy et al. 2017), we also allow for zero-inflation in the IA and RA processes by introducing intermediate Bernoulli filters: A prospect will only ever be acquired at all with probability $\pi^{(IA)}$, and after churning, a prospect will only ever be reacquired with probability $\pi^{(RA)}$. As in previous CBCV literature, we assume the size of the prospect pools are known in advance as they would not be separately identifiable from $\pi^{(IA)}$ and $\pi^{(RA)}$.

As discussed in Section 1, prior customer base analysis literature based on aggregated data has not separately modeled initial and repeat customer behavior. Incorporating granular panel data through data fusion allows us to separate out the two processes to understand how the behavioral patterns of repeat customers differ from those of first-time customers. We summarize the full individual-level model visually in Figure 1.

2.2. Parameter Heterogeneity

Consistent with prior literature, we incorporate unobserved heterogeneity by allowing the rate parameter λ in each process to vary across individuals.

Figure 1. Flow Diagram of the Proposed Individual-Level Acquisition and Retention Process



Notes. Weibull + PH(λ, c, $\boldsymbol{\beta}$) is shorthand notation for a proportional hazards model with a Weibull(λ, c) baseline hazard and $\boldsymbol{\beta}$ is a vector of covariate coefficients, discretized to the monthly level via differencing of the cumulative distribution function. The Bernoulli processes determine whether a customer is ever (re)acquired, whereas the respective Weibull processes determine time until (re)acquisition, given that (re)acquisition occurs.

Oftentimes a gamma heterogeneity distribution is used because of its conjugacy with the Weibull distribution (Schweidel et al. 2008b). In our model, we have a four-dimensional vector of rate parameters $\boldsymbol{\lambda}_i = (\lambda_i^{(IA)}, \lambda_i^{(IC)}, \lambda_i^{(RA)}, \lambda_i^{(RC)})'$. Although we could use independent gamma distributions for each parameter to retain conjugacy, we would also like to capture possible correlations between parameters: For instance, a positive correlation between $\lambda^{(IA)}$ and $\lambda^{(IC)}$ would indicate that early adopters also tend to be early abandoners, a pattern that cannot be captured by independent gamma distributions. Instead, we assume that each individual's rate parameter vector $\boldsymbol{\lambda}_i$ is drawn from a multivariate lognormal distribution[4]:

$$\log(\boldsymbol{\lambda}_i) \sim \mathcal{N}\big(\log(\boldsymbol{\lambda}_0), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\big). \tag{3}$$

2.3. Panel Selection

There may be selection bias that skews the distribution of behavioral patterns in the panel. We do not know the mechanism through which each individual i is selected into the panel, but without loss of generality, we can denote it by $P(Z_i = 1|\boldsymbol{\xi}_i) = f(\boldsymbol{\xi}_i)$ for some function f, where $\boldsymbol{\xi}_i$ is a vector of individual-level (possibly unobserved) characteristics on which individuals are selected into the panel. If the selection mechanism and the acquisition/churn processes are dependent (i.e., $\boldsymbol{\xi}_i \not\perp\!\!\!\perp \mathbf{Y}_i$, where $\mathbf{Y}_i$ is a random vector representing an individual i's acquisition and churn outcomes), then the selection process is nonignorable and must be corrected, or else it will skew parameter estimates and cause the inferences from the panel to not generalize to the population (Little and Rubin 2019). We model the selection mechanism jointly with $\mathbf{Y}_i$ to correct for this bias in the panel.

In our setting, it is reasonable to assume that $\boldsymbol{\xi}_i$ includes only ex ante heterogeneous characteristics, such that individuals are not directly selected into the credit card panel by their ex post behavior $\mathbf{Y}_i$; by definition, selection occurs before any granular behavior is observed in the panel. Hence, we assume the conditional independence $\boldsymbol{\xi}_i \perp\!\!\!\perp \mathbf{Y}_i|\boldsymbol{\lambda}_i$ holds. This assumption allows for the possibility that, for instance, wealthier customers may be more likely to be selected into a credit card panel (higher $P(Z_i = 1)$) and thus be more likely to sign up for Spotify (higher $\lambda_i^{(IA)}$), but assumes that their selection is not based on whether they *actually* sign up for Spotify ($\mathbf{Y}_i$). Thus, we model the probability that individual i is selected into the panel as a logistic regression on $\log(\boldsymbol{\lambda}_i)$:

$$\tilde{f}(\boldsymbol{\lambda}_i) := P(Z_i = 1|\boldsymbol{\lambda}_i) = Logit^{-1}\left(\beta_0^{(Z)} + \left(\beta^{(Z)}\right)'\log(\boldsymbol{\lambda}_i)\right). \tag{4}$$

Our assumed specification, $\tilde{f}(\boldsymbol{\lambda}_i)$, can be seen as a reduced-form approximation to the true selection mechanism $f(\boldsymbol{\xi}_i)$: $\boldsymbol{\xi}_i$ is unobserved but only introduces nonignorable selection bias via dependency with $\boldsymbol{\lambda}_i$; thus, modeling the selection mechanism through $\tilde{f}(\boldsymbol{\lambda}_i)$ controls for selection bias by indirectly controlling for dependency between $\boldsymbol{\xi}_i$ and $\mathbf{Y}_i$. The estimates of $\boldsymbol{\beta}^{(Z)}$ allow us to infer whether, for example, panel members are more or less prone to churning than members of the target population as a whole. In our setting, we do not have any observed covariates about panel members, but it would be straightforward to extend the selection function to also include data on observed characteristics.[5]

It is not immediately obvious that a model incorporating selection based on latent variables would be empirically identified. A key property in our context that allows for identification without needing to rely on distributional assumptions is that we have two sources of data about $\mathbf{Y}_i$: We have panel data, which may be contaminated by sample selection bias, and aggregate data, which covers the full target population. Intuitively, identifying the selection mechanism amounts to identifying the discrepancies between the panel and aggregate data in the periods they overlap. We formalize this intuition in Section 4.

Our approach for correcting for selection bias is analogous to the approaches used by Manchanda et al. (2004), Van Diepen et al. (2009), and Schweidel and

Knox (2013), who correct for nonrandom targeting of direct marketing by modeling targeting as a function of unobserved response heterogeneity. Schweidel and Moe (2014) use a similar approach to model consumer self-selection into posting on different online platforms.

The individual-level model, the heterogeneity distribution, and the panel selection mechanism jointly form our model specification; the full data generating process underlying our model specification is summarized in Online Appendix 1. The model parameters are $\boldsymbol{\lambda}_0$, $\boldsymbol{\Sigma}_\lambda$, $\beta_0^{(Z)}$, $\boldsymbol{\beta}^{(Z)}$, $\pi^{(IA)}$, $\pi^{(RA)}$, and $c^{(p)}$, $\boldsymbol{\beta}^{(p)}$ for $p \in \{IA, IC, RA, RC\}$.[6] We will refer to the concatenation of all of these parameters as $\boldsymbol{\theta}$.

3. Estimation Methodology

3.1. Observable Data

As discussed in Section 1.2, we do not observe all individual-level acquisition and churn data in our setting; instead, we observe some summary statistics aggregated across all individuals i and individual-level data for all panel members. Here, we formally define the observable panel data and aggregated summary statistics.

For this exposition, it is convenient to re-encode the acquisition and churn time outcomes into a series of binary random variables. Define $IA_{im} \in \{0, 1\}$ as a binary random variable equal to 1 if individual i is initially acquired in month m and takes value 0 otherwise. Define IC_{im}, RA_{im}, and RC_{im} analogously. Then, for a company that has been in commercial operations for M months, each individual's outcome up to month M can be represented as a $4M$-length vector of binary random variables, which we will call $\mathbf{Y}_i$:

$$\mathbf{Y}_i = (IA_{i1}, \ldots, IA_{iM}, IC_{i1}, \ldots, IC_{iM}, RA_{i1}, \ldots, RA_{iM}, RC_{i1}, \ldots, RC_{iM}). \tag{5}$$

First, we characterize the panel data. Through it, we implicitly observe for each individual a binary variable Z_i equal to 1 if individual i was selected into the panel and 0 if not. Conditional on individual i being in the panel, we observe a left-truncated version of $\mathbf{Y}_i$: that is, we observe all activity for individuals who are initially acquired after the panel is established, but do not observe any activity for individuals who are initially acquired before the panel is established. Denoting m^* as the starting month of panel data, the observable panel data, which we will call $\tilde{\mathbf{Y}}_i$, for an individual i who is in the panel, is as follows:

$$\left(\tilde{\mathbf{Y}}_i | Z_i = 1\right) = \left(IA_{im^*}, \ldots, IA_{iM}, IC_{im^*}, \ldots, IC_{iM}, RA_{im^*}, \ldots, RA_{iM}, RC_{im^*}, \ldots, RC_{iM}\right). \tag{6}$$

As such, the length of the observation period in the panel data are $M - m^{*+1}$ months. If individual i was not selected into the panel, then we observe no panel data for that individual: that is, $(\tilde{\mathbf{Y}}_i | Z_i = 0) = \emptyset$. In our empirical context, observations begin for all credit card panel members at the same time, but this exposition could easily be generalized to imbalanced panels where m^* is individual-specific.

Next, we characterize the aggregate data. The three summary statistics disclosed by Spotify (hereafter referred to interchangeably as *disclosures*) in our empirical application are the gross number of subscribers added and lost in quarter q (which we call ADD_q and $LOSS_q$, respectively), and the total count of active subscribers at the end of quarter q (END_q). These summary statistics are also the most commonly disclosed by publicly traded subscription-based companies (McCarthy et al. 2017) and can be expressed as aggregations of linear combinations of the elements of $\mathbf{Y}_i$:

$$\begin{aligned} ADD_q &= \sum_{i=1}^{N} \sum_{m=3q-2}^{3q} IA_{im} + RA_{im} \\ LOSS_q &= \sum_{i=1}^{N} \sum_{m=3q-2}^{3q} IC_{im} + RC_{im} \\ END_q &= \sum_{q^*=1}^{q} ADD_{q^*} - LOSS_{q^*}, \end{aligned} \tag{7}$$

where N is the total size of the target population. The random vector of aggregate data observations, which we will call $\mathbf{D}_N$, is the concatenation of all observed values of ADD_q, $LOSS_q$, and END_q.

3.2. The Proxy Likelihood Function

A seemingly natural approach to estimate our model with this observable data would be to use likelihood-based methods, such as maximum likelihood estimation. The full log-likelihood of all observed data can be decomposed as follows:

$$\begin{aligned} \ell(\boldsymbol{\theta} | z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}) = &\sum_{i=1}^{N} \underbrace{\log(P_{\boldsymbol{\theta}}(Z_i = z_i))}_{\text{selection outcomes}} \\ &+ \sum_{\{i | z_i = 1\}} \underbrace{\log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i | Z_i = z_i\right)\right)}_{\text{panel data}} \\ &+ \underbrace{\log\left(P_{\theta}\left(\begin{matrix}\mathbf{D}_N = \mathbf{d} | Z_{1:N} = z_{1:N}, \\ \tilde{\mathbf{Y}}_{1:N} = \tilde{\mathbf{y}}_{1:N}\end{matrix}\right)\right)}_{\text{aggregate data}}. \end{aligned} \tag{8}$$

Selection bias is accounted for in the panel log-likelihood and aggregate data log-likelihood by conditioning on Z_i in the second and third terms of this equation, and the aggregate data avoids double-counting the panel data by conditioning on $\tilde{\mathbf{Y}}_i$.

Conditional upon the individual-level vectors $\boldsymbol{\lambda}_{1:N}$, the first two terms in Equation (8) are simple to compute: The first term is the panel selection likelihood, which is the likelihood of a logistic regression model; the second term is the panel data likelihood, which consists of duration probabilities computed directly from our model. The third term, however, requires an N-fold convolution over $\mathbf{Y}_{1:N}$ and is not tractable to compute. As such, typical likelihood-based estimators are infeasible. Alternately, we could consider simply using moment-based methods such as nonlinear least squares, which are often relatively simple to implement and have been used extensively in prior CBCV literature (Gupta et al. 2004, McCarthy et al. 2017, McCarthy and Fader 2018). However, this would require summarizing the granular data $\{Z_{1:N}, \tilde{\mathbf{Y}}_{1:N}\}$ into aggregate moments, while only relatively simple models admit sufficient statistics that allow summarization without information loss. This is the approach taken by Berry et al. (2004), but it requires the potentially arbitrary choice of which moments to include, possibly hurting statistical efficiency. In our setting, it is unclear what summary statistics would adequately summarize the panel data with minimal information loss. Instead, our proposed estimation procedure incorporates both the aggregate and panel data as-is, without information loss; this sharpens our model estimates and improves the portability of our method to other domains in which the disaggregate data may be difficult to aggregate into summary statistics.

To the best of our knowledge, the primary prior work that has proposed an estimator applicable to our setting is that of Feit et al. (2013), who use Bayesian imputation, building off prior work on Bayesian estimation of individual-level models of consumer choice using only aggregated data (Chen and Yang 2007; Musalem et al. 2008, 2009). In the Bayesian imputation approach, the missing observations in $\mathbf{Y}_{1:N}$ are treated as parameters to be estimated along with all model parameters; augmented data sets $\hat{\mathbf{Y}}_{1:N}$ representing possible values of $\mathbf{Y}_{1:N}$ are simulated, with proposed augmented data sets accepted only if the resulting aggregate data points $\hat{\mathbf{D}}_N$ are equal to the observed aggregate data points $\mathbf{d}$.

This method performs well when the overall population N is not large. However, it is computationally infeasible to scale to large-data settings such as ours: A $\mathbf{Y}_i$ vector must be imputed for all N population members, and when the vector of summary statistics $\mathbf{d}$ is high-dimensional, there will be a large number of equality constraints that proposed data sets $\hat{\mathbf{Y}}_{1:N}$ will be unlikely to satisfy, resulting in low acceptance ratios and thus poor mixing and slow convergence. In our empirical application, N is six orders of magnitude larger than in Feit et al. (2013), making this approach infeasible.

To ameliorate the scalability issues, subsampling the data (e.g., scaling down the population size and aggregate count data by a multiplicative factor) has been proposed in previous work (Musalem et al. 2009). This approach, although getting around scalability issues, results in poor statistical efficiency, because subsampling the data inflates estimation variance. For instance, scaling down the aggregate data by 1,000 times would result in standard errors that are $\sqrt{1{,}000} \approx$ 31.6 times larger than when using the full data.

The approaches of Berry et al. (2004) and Feit et al. (2013) could be modified to correct for selection bias by computing moments and probabilities conditional on panel selection outcomes as in Equation (8). However, because of the previous issues of summarizing the panel data (in the former case) and scalability (in the latter case), we instead propose model estimation by maximizing a *proxy likelihood* function, which replaces the third term in Equation (8) by a computationally tractable approximation. Recall that this term is the log-likelihood of $\mathbf{D}_N$, which is the aggregation of (a linear transformation of) N individual-level outcomes $\mathbf{Y}_i$. Thus, under mild regularity conditions, the central limit theorem states that the distribution of $\mathbf{D}_N$ is asymptotically well approximated by a normal distribution. Hence, we approximate the likelihood of $\mathbf{D}_N$ using its asymptotic distribution.[7]

In particular, define the finite sample conditional mean and variance of $\mathbf{D}_N$ as follows:

$$\begin{aligned} \boldsymbol{\mu}_N(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}\left[\mathbf{D}_N | z_{1:N}, \tilde{\mathbf{y}}_{1:N}\right] \\ \boldsymbol{\Sigma}_N(\boldsymbol{\theta}) &= \mathrm{Var}_{\boldsymbol{\theta}}\left[\mathbf{D}_N | z_{1:N}, \tilde{\mathbf{y}}_{1:N}\right]. \end{aligned} \tag{9}$$

The distribution $MVN(\boldsymbol{\mu}_N(\boldsymbol{\theta}), \boldsymbol{\Sigma}_N(\boldsymbol{\theta}))$ is an asymptotic approximation to the distribution of $\mathbf{D}_N$ given the panel data, so we can approximate the true likelihood function from Equation (8) by replacing the last term with the log-density of this multivariate normal distribution. Thus, we have replaced the computationally prohibitive task of computing the full distribution of $\mathbf{D}_N$ by the much more manageable task of computing its first two moments, similar to what would be required for other moment-based procedures such as two-stage generalized method of moments (Hansen 1982). A typical moment-based estimator would require only the unconditional moments of $\mathbf{D}_N$, whereas here we condition on the disaggregate data to avoid double-counting the panel members (i.e., computing the likelihood of the panel members directly through their panel activity and again indirectly through their representation within the aggregate data); however, we could easily replace $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ with their unconditional analogues without fundamentally altering the properties of our estimator when the conditional moments are infeasible to compute.

The second moment $\boldsymbol{\Sigma}_N$ may be computationally expensive to obtain, because the number of covariance elements to compute will be very large when $\mathbf{D}_N$

is high-dimensional.[8] In contrast, the first moment $\boldsymbol{\mu}_N(\boldsymbol{\theta})$ is relatively easy to compute, because it scales linearly with the dimension of $\mathbf{D}_N$ (and would also be required for any typical moment-based estimator such as nonlinear least squares). As such, we consider an estimator where only $\boldsymbol{\mu}_N(\boldsymbol{\theta})$ is computed in the optimization, and the covariance matrix is fixed at some positive definite matrix $\hat{\boldsymbol{\Sigma}}_N$ instead of being continuously updated (we will discuss in Section 3.4 the matter of choosing an appropriate matrix $\hat{\boldsymbol{\Sigma}}_N$). That is, dropping the conditioning on $z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}$ for notational brevity, the proxy likelihood $\tilde{\ell}$ is as follows (up to additive constants):

$$\begin{aligned}\tilde{\ell}_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right) = &\sum_{i=1}^{N} \log(P_{\boldsymbol{\theta}}(Z_i = z_i)) \\ &+ \sum_{\{i|z_i=1\}} \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i|Z_i = z_i\right)\right) \\ &- \frac{1}{2}\left(\mathbf{d} - \boldsymbol{\mu}_N(\boldsymbol{\theta})\right)'\left(\hat{\boldsymbol{\Sigma}}_N\right)^{-1}\left(\mathbf{d} - \boldsymbol{\mu}_N(\boldsymbol{\theta})\right). \qquad (10)\end{aligned}$$

The third term in Equation (10) is the multivariate normal approximation to the log-likelihood of $\mathbf{d}$. We can also see this third term as a quadratic form of moment conditions, which is equivalent to the objective function of the generalized method of moments, up to normalization by N (Hansen 1982). Equation (10) as a whole is the objective function of our estimation procedure. In Section 3.3, we describe how to compute the proxy likelihood for our empirical specification and then discuss our proposed estimation procedure in Section 3.4.

3.3. Computing the Proxy Likelihood

Although we have reduced the computational problem of the aggregate data likelihood down to computing the first two moments of the aggregate summary statistics as defined in Equation (9), there remains the issue of how to compute these moments. As noted in Section 3.1, all summary statistics we observe are simply aggregations of linear transformations of $\mathbf{Y}_i$, which makes it straightforward to compute these quantities as a function of the conditional mean vector and covariance matrix of $\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i$. Thus, we just need to compute the moments of $\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i$ to derive the moments of the summary statistics.[9]

Because $\mathbf{Y}_i$ is a vector of Bernoulli random variables, computing its first two moments requires computing all marginal and pairwise event probabilities. These probabilities are not straightforward to compute, because a single customer may engage in repeat behaviors (adoption and/or churn) more than once: For instance, naively computing probabilities such as $P(RA_{im} = 1)$ would require marginalizing over all possible sequences of acquisitions and churns that the customer could have taken to arrive at reacquisition in month m, which is computationally prohibitive even for short time horizons. To make the marginalization feasible, we construct a recursive belief propagation algorithm that exploits the Markovian structure of our model, substantially reducing computational burden (Pearl 1988).

Finally, we must also marginalize the individual-level parameters $\boldsymbol{\lambda}_{1:N}$ out of the objective function. As mentioned in Section 2.2, our heterogeneity distribution is nonconjugate; as such, we approximate the integrals numerically via simulation using Halton sequences, which have been used successfully in empirical applications similar to ours (Bhat 2001, Train 2009). In essence, this amounts to simulating K draws from the mixing distribution $\boldsymbol{\lambda}^{(k)} \sim g(\boldsymbol{\lambda})$ using a Halton sequence, computing each probability and expectation in Equation (10) conditional upon $\boldsymbol{\lambda}^{(k)}$, and then averaging over the K draws before plugging the results into the proxy likelihood equation. Because the second and third terms of Equation (10) involve probabilities and expectations conditional upon the panel selection outcome Z_i and the panel data $\tilde{\mathbf{Y}}_i$, for these terms, we integrate over the posteriors $g(\boldsymbol{\lambda}_i|Z_i)$ and $g(\boldsymbol{\lambda}_i|Z_i, \tilde{\mathbf{Y}}_i)$ using importance sampling, weighting each $\boldsymbol{\lambda}^{(k)}$ by the probability of the data being conditioned on. Pseudocode and a step-by-step procedure for how to compute the moments $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$, including a derivation of the aforementioned belief propagation algorithm, are provided in Online Appendix 2.

These computational ingredients provide all the tools needed to efficiently compute the proxy likelihood $\tilde{\ell}$, such that we can use our method in practice. The full procedure for computing $\tilde{\ell}$ is summarized in Algorithm 1. With our objective function in hand, we now discuss our estimation procedure.

Algorithm 1 Pseudocode for Computing Objective Function $\tilde{\ell}$

```
function ProxyLL($\boldsymbol{\theta}, z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}, \hat{\boldsymbol{\Sigma}}_N, K$)
    Simulate $K$ draws of $\boldsymbol{\lambda}$ and compute conditional panel selection probabilities for each draw:
    for all $k = 1, 2, \ldots, K$ do
        Using the $k$-th term of a four-dimensional Halton sequence, simulate $\boldsymbol{\lambda}^{(k)}$ from the mixing distribution
            $\log(\boldsymbol{\lambda}^{(k)}) \sim \mathcal{N}(\log(\boldsymbol{\lambda}_0), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}})$
        Compute the probability of selection into the panel given $\boldsymbol{\lambda}^{(k)}$ (Equation (4)):
            $p_k := P_\theta(Z = 1|\boldsymbol{\lambda}^{(k)})$
                $= Logit^{-1}\left(\beta_0^{(Z)} + (\boldsymbol{\beta}^{(Z)})' \log(\boldsymbol{\lambda}^{(k)})\right)$
```

Compute the panel selection log-likelihood (first term of Equation (10)) by averaging over the p_ks:

$$\ell_z := \sum_{i=1}^{N} z_i \log\left(\frac{1}{K}\sum_{k=1}^{K} p_k\right) + (1-z_i)\log\left(1-\frac{1}{K}\sum_{k=1}^{K} p_k\right).$$

Compute the panel data log-likelihood (second term of Equation (10)) by averaging each panel member's likelihood over the simulated $\boldsymbol{\lambda}^{(k)}$s and summing the marginal log-likelihood over panel members, where the p_ks serve as posterior importance sampling weights, as described in Online Appendix 2.3:

$$\ell_y := \sum_{\{i|z_i=1\}} \log\left(\frac{1}{\sum_{k=1}^{K} p_k}\sum_{k=1}^{K} p_k P_\theta\left(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i \middle| Z_i = z_i, \boldsymbol{\lambda}_i = \boldsymbol{\lambda}^{(k)}\right)\right).$$

Compute the aggregate proxy likelihood (third term of Equation (10)) by first computing the mean function $\boldsymbol{\mu}_N(\boldsymbol{\theta})$ using Algorithm 2 in the online appendix and then plugging this into the multivariate normal log-density formula for aggregate data $\mathbf{d}$, yielding (up to additive constants):

$$\ell_d := -\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}_N(\boldsymbol{\theta}))'(\hat{\boldsymbol{\Sigma}}_N)^{-1}(\mathbf{d}-\boldsymbol{\mu}_N(\boldsymbol{\theta})).$$

Sum the three terms to compute the total proxy log-likelihood for the parameters $\boldsymbol{\theta}$:

$$\tilde{\ell} := \ell_z + \ell_y + \ell_d$$

return $\tilde{\ell}$

3.4. Maximum Proxy Likelihood Estimation

Section 3.2 introduced the proxy likelihood function $\tilde{\ell}$ and Section 3.3 described how to compute it. With this objective function, we can construct a one-stage estimator as follows:

$$\hat{\boldsymbol{\theta}}_N^{(1)}(z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}|\hat{\boldsymbol{\Sigma}}_N) = \arg\max_{\boldsymbol{\theta}} \tilde{\ell}_N(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N)$$

for some $q \times q$ positive definite matrix $\hat{\boldsymbol{\Sigma}}_N$, where q is the dimension of $\mathbf{d}$. However, we must also consider the practical matter of choosing the covariance matrix $\hat{\boldsymbol{\Sigma}}_N$ to ensure that the normal approximation is accurate. As such, we propose a two-stage estimator $\hat{\boldsymbol{\theta}}_N^{(2)}(z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}|\hat{\boldsymbol{\Sigma}}_N)$, which updates $\hat{\boldsymbol{\Sigma}}_N$, analogous to the two-stage generalized method of moments procedure for weight matrix selection (Hansen 1982):

1. Initialize $\hat{\boldsymbol{\Sigma}}_N$ to some $q \times q$ positive-definite matrix (e.g., the identity matrix or the true covariance matrix at a heuristic estimate of $\boldsymbol{\theta}$).

2. Obtain an initial parameter estimate $\tilde{\boldsymbol{\theta}}$ by maximizing the proxy likelihood given $\hat{\boldsymbol{\Sigma}}_N$:

$$\tilde{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}_N^{(1)}(z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}|\hat{\boldsymbol{\Sigma}}_N).$$

3. Update covariance matrix $\hat{\boldsymbol{\Sigma}}_N$ to the true covariance matrix at the initial estimate $\tilde{\boldsymbol{\theta}}$ (Algorithm 3 in the online appendix):

$$\hat{\boldsymbol{\Sigma}}_N \leftarrow \boldsymbol{\Sigma}_N(\tilde{\boldsymbol{\theta}}).$$

4. Update $\tilde{\boldsymbol{\theta}}$ using the updated covariance matrix $\hat{\boldsymbol{\Sigma}}_N$:

$$\tilde{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}_N^{(1)}(z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}|\hat{\boldsymbol{\Sigma}}_N).$$

5. Return the updated $\tilde{\boldsymbol{\theta}}$ as the final parameter estimate $\hat{\boldsymbol{\theta}}_N^{(2)}$.

Under mild regularity conditions analogous to those of maximum likelihood and generalized method of moments, $\hat{\boldsymbol{\theta}}_N^{(1)}$ and $\hat{\boldsymbol{\theta}}_N^{(2)}$ are consistent, converging at rate $O_p(N^{-1/2})$ and achieving asymptotic normality (derivations are available in Online Appendix 3).[10] We will refer to these estimation methods collectively as maximum proxy likelihood (MPL). Although we use the two-stage estimator $\hat{\boldsymbol{\theta}}_N^{(2)}$ in all our simulations and in our empirical example for its statistical efficiency, the single-stage estimator $\hat{\boldsymbol{\theta}}_N^{(1)}$ may still be useful for models where computing the second moment $\boldsymbol{\Sigma}_N(\boldsymbol{\theta})$ is infeasible. To calculate the standard errors associated with this estimate, we derive its asymptotic variance in Online Appendix 3.[11]

We now have a computationally feasible and statistically efficient method for estimating our model. In the following sections, we demonstrate the validity of our method in several ways: first through a discussion of model identification in Section 4, then through a simulation study in Section 5, and finally through our empirical example of Spotify in Section 6.

4. Identification

We discuss how changes in the model parameters produce distinct patterns in the aggregate and disaggregate data, identifying the model. We first discuss identification of all parameters that are homogeneous across customers and then the parameters that govern cross-sectional heterogeneity in customers' propensities, before concluding with an analogous discussion of the parameters and functional form of the selection model.

4.1. Homogeneous Parameters

The homogeneous parameters of our specification are the process-specific covariate coefficients $\boldsymbol{\beta}^{(p)}$ and duration dependence parameter $c^{(p)}$ for the four processes $p \in \{IA, IC, RA, RC\}$, as well as the proportion of the population who will ever be acquired $\pi^{(IA)}$, and the proportion of churned customers who will ever be reacquired, $\pi^{(RA)}$.

First, we consider $\boldsymbol{\beta}^{(p)}$. In our empirical application, $\mathbf{x}_t$ is a vector of quarterly dummies. In our panel data, we can directly observe quarterly seasonality in initial and repeat acquisition and churn behaviors, separately identifying seasonality for the four processes. These parameters are further identified through the aggregate data by quarterly acquisitions and churns (ADD_q and $LOSS_q$, respectively). Although our aggregate data do not distinguish between first-time and repeat activity, they can still facilitate separate identification of initial and repeat behaviors through changes in seasonality over time: As a company matures, an increasing proportion of its acquisitions and churns will be attributable to repeat behaviors; therefore, long-term trends in the strength of seasonality distinguish between initial and repeat processes. The effects of other time-varying covariates can be identified similarly.

Identification of the duration dependence parameters $c^{(p)}$ follows from observing long-term trends in acquisition and churn counts in the panel data, which trace out the baseline shape of the empirical hazard function for each process. Unobserved heterogeneity in the scale parameter $\lambda_i^{(p)}$ results in a decreasing hazard function because of survivorship bias, which can be difficult to separate out from duration dependence. In general, duration dependence and heterogeneity can only be separately identified under restrictions on the separatibility and/or parametric form of the hazard function (Heckman 1991). These conditions are satisfied under our model, enabling identification; that being said, identification may be sensitive to violations of model assumptions that are difficult to verify, and some degree of misspecification is inevitable in practice. As such, although our model is formally identifiable, we should nevertheless interpret the duration dependence estimates with some caution.

The parameters $\pi^{(IA)}$ and $\pi^{(RA)}$ are identified by observing the asymptote of where the initial and repeat acquisition curves flatten out: For instance, in the panel data, we can directly observe the proportion of panel members who have been initially acquired and the proportion of previously churned panel members who have been reacquired. With a long enough panel horizon, we can infer the asymptotes of the cumulative acquisition curves that determine $\pi^{(IA)}$ and $\pi^{(RA)}$. Identification of these parameters will naturally be weaker for companies that are still rapidly growing or have very long acquisition cycles, because it will be more difficult to infer the asymptotes.

4.2. Heterogeneity Distribution Parameters

The parameters governing the heterogeneity distribution of $\boldsymbol{\lambda}_i$ are the log-mean vector $\boldsymbol{\lambda}_0$ and covariance matrix $\boldsymbol{\Sigma}_\lambda$. We first discuss how the aggregate data identifies the log-mean and variance of $\boldsymbol{\lambda}_i$. We then discuss how the panel data identifies the distribution of $\boldsymbol{\lambda}_i$, up to distortion by selection bias.

Although the acquisition and retention curves for each process are not directly observable in the aggregate data, ADD_q and $LOSS_q$ are nevertheless informative about the distribution of $\boldsymbol{\lambda}_i$. The log-means of the distributions, $\boldsymbol{\lambda}_0$, are readily identified by the scale of the ADD_q and $LOSS_q$ curves: Earlier periods, in which most customers are first-timers, identify $\lambda_0^{(IA)}$ and $\lambda_0^{(IC)}$; later periods, in which most customers are repeaters, identify $\lambda_0^{(RA)}$ and $\lambda_0^{(RC)}$.

The variance of $\lambda_i^{(IA)}$ is identified by the shape of the ADD_q curve in early periods: Low variance suggests near-exponentially declining curves (augmented by duration dependence), whereas high variance suggests a steep initial drop off followed by a quick flattening out because of survivorship bias. The distribution of $\lambda_i^{(IC)}$ is identified by observing sequential correlations between the $LOSS_q$ and the ADD_q curves: For instance, if in the quarter after a spike of high acquisitions (e.g., because of seasonality), there is a corresponding spike in $LOSS_q$, this suggests high heterogeneity variance; conversely, if $LOSS_q$ does not spike sharply immediately after spikes in ADD_q, this suggests low heterogeneity variance. The variance of $\lambda_i^{(RA)}$ and $\lambda_i^{(RC)}$ are identified through analogous patterns in later periods.

Next, we consider identification of the heterogeneity parameters through the panel data. The following arguments imply identification of the heterogeneity distribution of the *panel members* and not the *population as a whole*. As we will discuss in the next section, this distinction will be key to identifying the selection parameters.

The panel-level marginal distribution of $\lambda_i^{(p)}$ for each process is identified from the shape of the initial/repeat acquisition and retention curves, analogous to identification of the distribution of $\lambda_i^{(IA)}$ from the aggregate ADD_q curve.

Additionally, the correlations between the different process $\lambda_i^{(p)}$s (i.e., the off-diagonal elements of $\boldsymbol{\Sigma}_\lambda$) are identified directly by joint observations in the panel data: For instance, if panel members who are first acquired early on tend to churn more quickly than panel members acquired later on, this suggests a positive correlation between $\lambda_i^{(IA)}$ and $\lambda_i^{(IC)}$ (Schweidel et al. 2008a). It is less obvious how such correlations would manifest in the aggregate data, where we cannot observe the joint distribution of acquisition and churn times. For these parameters, the panel data are most informative.

Although we use a lognormal heterogeneity specification, a similar identification argument applies to other distributional forms: The presence of high $\lambda_i^{(p)}$s reflects mostly at the beginning of a process, whereas the presence of low $\lambda_i^{(p)}$s reflects later on, and dependencies between the different processes can be inferred directly by the joint distribution of acquisition and churn times in the panel data.

4.3. Selection Parameters and Functional Form

The identification of the selection function intuitively boils down to observing discrepancies between the acquisition and churn trends in the panel data relative to the aggregate data. We argued previously that the population mean of $\boldsymbol{\lambda}_i$ is readily identified in the aggregate data, whereas the mean of $\boldsymbol{\lambda}_i$ among panel members is identified through the panel data; accordingly, first-order selection bias is identified by comparing the means implied by the aggregate data versus the panel data. For instance, a positive $\beta_{IA}^{(Z)}$ means that the panel population has higher $\lambda_i^{(IA)}$ on average compared with the population as a whole, such that the acquisition rates observed in the panel in early periods will be higher than the acquisition rates implied by the aggregate ADD_q figures. Analogously, $\beta_{IC}^{(Z)}$, $\beta_{RA}^{(Z)}$, and $\beta_{RC}^{(Z)}$ are identified by discrepancies between panel and aggregate moments that identify the means of the corresponding $\lambda_i^{(p)}$ distributions. Finally, the intercept term $\beta_0^{(Z)}$ simply governs the size of the panel and is identified by the empirical proportion of the total population that is in the panel.[12]

Selection model specifications other than a logit-linear form can also capture these first-order differences, but the proposed functional form should be a reasonable approximation to smooth, monotonic selection functions, such that bias from misspecification of the functional form should be small. We verify this through simulation studies, available in Online Appendix 5.4, where the true data generating process has a nonlinear selection function, whereas the estimated model assumes a linear selection function. We find that bias and coverage worsen, but the method still mostly recovers the true population parameters.

Although we use a simple linear specification for our selection function for parsimony, we can also specify more sophisticated nonlinear selection mechanisms that could capture second-order selection effects such as nonmonotonicity. In addition to the mean vector $\boldsymbol{\lambda}_0$, we argued that the overall marginal heterogeneity distributions are identified through the aggregate data. Thus, nonlinearities in the selection function can be identified by discrepancies between the moments that identify different quantiles of the distributions.[13]

Formally, we show in Online Appendix 4 that as long as the panel data are sufficiently rich, a general selection mechanism $P_{\psi}(Z_i = 1|\boldsymbol{\lambda}_i)$, parameterized by vector $\boldsymbol{\psi}$, is identified as long as there are enough aggregate moments to separate out the discrepancies between the overall population and the panel population implied by each element of $\boldsymbol{\psi}$; at minimum, there must be as many unique aggregate moments as there are elements of the selection parameter $\boldsymbol{\psi}$. The intuition for this result is that if the panel data are rich, we can identify all the parameters of the model except $\boldsymbol{\psi}$ on the panel data alone, leveraging all of the aggregate data to estimate $\boldsymbol{\psi}$ by observing the discrepancies between the trends in the aggregate data and those implied by the parameters identified by the panel data.

In practice, however, there is a bias-variance tradeoff to consider: Allowing for a more flexible selection model reduces bias from misspecification but inflates variance by adding more parameters to estimate. In particular, second-order selection biases are likely to be imprecisely estimated, because the moments needed to identify them will be relatively noisy compared with those needed to identify first-order selection bias.[14] Unless we have strong reason to believe that misspecification because of second-order selection biases is severe, a linear selection model is likely to perform better, because it can capture first-order shifts between the panel and aggregate data but can still be relatively precisely estimated. In our empirical application, incorporating panel data with a linear selection mechanism yields better out-of-sample predictions of future aggregate data than using historical aggregate data alone. Although this is in no way a guarantee that our model does not have misspecification, it demonstrates that the linear selection mechanism corrects for selection bias well enough to make the panel data useful in capturing and forecasting population-level trends.

5. Simulation Studies

5.1. Predictive Accuracy

We conduct simulation studies to evaluate the performance of the MPL method, comparing it with two other methods commonly used in practice: First, the generalized method of moments with only aggregate data, as in all extant CBCV models (denoted AGG); and second, maximum likelihood with only granular panel data, assuming that there is no selection bias (denoted PAN). We note that MPL requires about as much computing time as AGG and PAN combined and therefore has small incremental cost of implementation in terms of computation. Compute times by method are reported in Online Appendix 5.2.

We compare AGG and PAN to MPL across a variety of simulation settings to better understand the incremental improvement our method provides as a function of contextual factors. We vary these settings along two groups of dimensions:

- Data settings: Our baseline data setting has $M = 60$, $N = 100K$, panel size equal to 5% of N, and a moderate degree of selection bias (through $\boldsymbol{\beta}^{(Z)}$). We then perturb this baseline scenario marginally, considering perturbations (1) $M \in \{36, 84, 108\}$, (2) $N \in \{20K, 500K, 2.5M\}$, (3) panel percentages of 1% and 10%, and (4) selection bias severities of none and high. This results in 11 total data settings.

- Parameter settings: We separately vary the model parameters governing the initial and repeat acquisition and churn processes along eight dimensions: (1–4) the baseline parameters (λ_0, c) for these processes, (5 and 6) initial and repeat process variance, and (7 and 8) within- and across-process correlation. We consider low and high values for each dimension that is varied, resulting in $2^8 = 256$ unique sets of parameter values.

This results in a total of 2,816 simulation settings (see Online Appendix 5.1 for the complete listing of settings). For consistency with the data available in our Spotify application, we assume that *END* and *LOSS* data are observed each quarter (which fully determine ADD_q, because $ADD_q = END_q - END_{q-1} + LOSS_q$). In this section and the next, we evaluate the methods based on predictive accuracy and parameter recovery, respectively, to establish the usefulness of the method for both prediction-oriented use cases (e.g., CBCV) and inference-oriented ones.

We evaluate predictive accuracy in this section by forecasting quarterly initial and repeat acquisition and churn (*QIA*, *QIC*, *QRA*, *QRC*), in addition to *ADD*, *LOSS*, and *END*, over a six-quarter holdout period. The former collection of summary statistics separates out initial and repeat behavior, whereas the latter collection pools them. Our error measure is mean absolute percentage error (MAPE) so that error measures are comparable across the disclosures (i.e., summary statistics) despite their differing scales. The true values underlying the MAPE calculations are the actual values of the aggregate statistics in the holdout period. To better understand the overall performance of each method, Table 1 shows the MAPE for each method by disclosure, averaging across all parameter and data settings. MPL has the lowest average MAPE figures across all disclosures. Its improvements over the other methods are particularly sizable when predicting the summary statistics that separate out initial and repeat behavior (*QIA*, *QIC*, *QRA*, *QRC*), because it could better infer these disclosures through the panel data. Its relative advantage remains substantial for *ADD*, *LOSS*, and *END* as well.[15] The accuracy of AGG is generally high when predicting disclosures that are observed historically, deteriorating significantly for disclosures that are not. PAN has low accuracy in general, further emphasizing the perils of naively using granular data when it may not be representative of the target population.

It also important to understand how performance varies as a function of the characteristics of the available data. Figure 2 plots average MAPE figures by method as we vary the four data settings. These figures are averaged across *QIA*, *QIC*, *QRA*, and *QRC*. Disclosure-specific MAPE data are available in Online Appendix 5.3. MPL consistently outperforms PAN and AGG. The performance gap narrows, particularly for AGG, as *N* increases, and the gap for PAN narrows as the panel selection bias decreases and the panel size increases. PAN performs worst, except when there is no selection bias.

Table 1. Simulation Study: Holdout MAPE by Method and Disclosure, Averaging Across Settings

Disclosure	PAN	AGG	MPL
QIA	31.0%	23.5%	9.2%
QIC	10.6%	23.0%	8.0%
QRA	17.1%	9.1%	3.9%
QRC	17.2%	10.3%	4.2%
ADD	17.8%	2.4%	2.0%
LOSS	14.5%	2.6%	2.1%
END	84.6%	0.9%	0.5%

5.2. Parameter Recovery

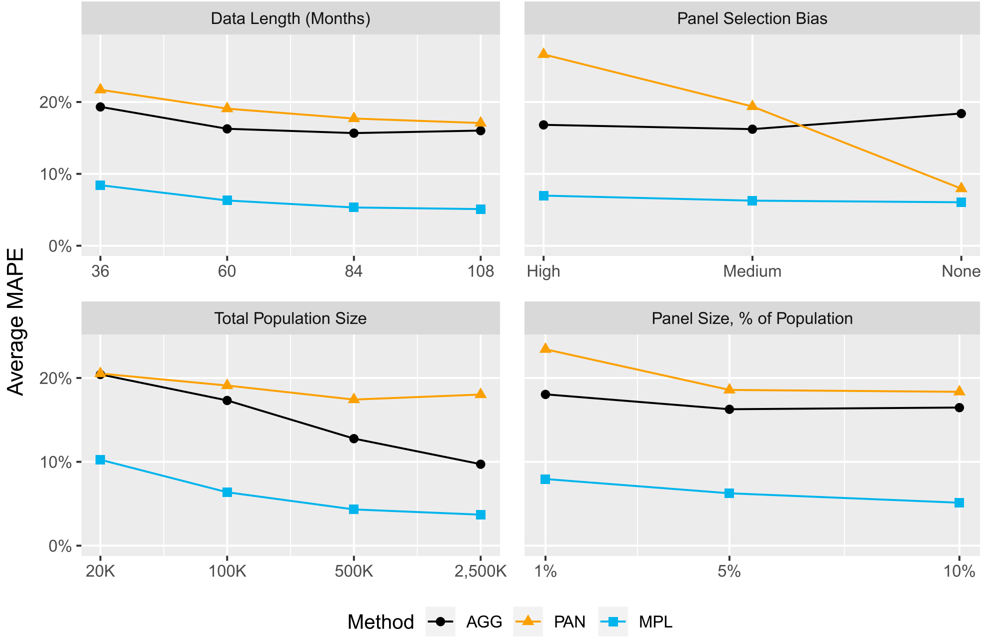
Although predictive accuracy is important for our empirical application, it is also important to evaluate parameter recovery, particularly for other settings where inference may be a primary objective. To this end, we conduct simulations in this section to evaluate MPL's finite sample parameter recovery performance.

We evaluate MPL using the baseline set of parameter values from the preceding large-scale simulation analysis.[16] As a robustness check, we repeat the analysis for another four randomly selected parameter sets from the previous section (the results, which are qualitatively consistent with the results reported here, are provided in Online Appendix 5.4).

First, we compare the bias and variance of MPL to AGG by computing the median absolute bias and interquartile range (IQR) for each parameter, averaged across 30 replicates. For brevity, we grouped the model parameters into five categories, heterogeneity mean parameters ($\lambda^{(IA)}, \lambda^{(IC)}, \lambda^{(RA)}, \lambda^{(RC)}$), heterogeneity variance parameters ($\sigma_\lambda^{(IA)}, \ldots, \sigma_\lambda^{(RC)}$), heterogeneity correlation parameters ($\rho_\lambda^{(IA,RA)}, \ldots, \rho_\lambda^{(RA,RC)}$), homogeneous parameters ($\pi^{(IA)}, \pi^{(RA)}, c^{(IA)}, \ldots, c^{(RC)}$), and selection parameters ($\beta_0^{(Z)}$, $\boldsymbol{\beta}^{(Z)}$), and report each performance measure averaged across all parameters within each category (parameter-by-parameter results are available in Online Appendix 5.4). We compute each parameter's statistics in terms of absolute percentage to account for differing scales and signs of parameters. For example, the median absolute percentage bias for parameter collection c is equal to

$$\text{MAPB}(c) = \sum_{p=1}^{n_p^c} \frac{\left|\text{Med}\left(\hat{\boldsymbol{\theta}}_{c(p)}\right) - \boldsymbol{\theta}_{c(p)}\right|}{\left|\boldsymbol{\theta}_{c(p)}\right|},$$

Figure 2. (Color online) Simulation Study: Holdout MAPE for *QIA*, *QIC*, *QRA*, and *QRC* by Method and Data Setting, Averaging across Parameter Settings and Disclosures



where n_p^c is the number of parameters within parameter collection c, $\boldsymbol{\theta}_{c(p)}$ denotes the true value of the pth parameter within parameter collection c, and $\text{Med}(\hat{\boldsymbol{\theta}}_{c(p)})$ represents the sample median of parameter estimate $\hat{\boldsymbol{\theta}}_{c(p)}$ across simulation replicates. We use median and IQR to be robust to outliers, as we found the AGG method yielded very heavy-tailed estimate distributions. The results are shown in Table 2.

Median bias and IQR figures are generally good for MPL for each parameter category: Bias is low and the IQR is generally small. In contrast, median bias and IQR are one to three orders of magnitude larger for AGG than for MPL within each parameter category. Although AGG is asymptotically consistent, it is evidently not empirically identifiable with five years of quarterly data summaries. These results are not sensitive to the true parameter values we select: AGG fails similarly in other parameter settings.

Next, we focus our study on MPL using more traditional performance measures. In Table 3, we compute the mean absolute bias, coefficient of variation of estimates, and empirical coverage rate of a 95% confidence interval (using the asymptotic variance formula derived in Online Appendix 3 to calculate standard errors). Table 3 suggests that, under correct specification, MPL's parameter recovery performance is good. Mean absolute bias as a percentage of true parameter values is less than 7% for all parameter categories. The coefficient of variation (CV) of the parameter estimates was less than 11% across all parameter categories except the heterogeneity correlation parameters, for which the CV was 42%. Empirical coverage is equal to its theoretical target level within simulation error. In Online Appendix 5.4, we also report results when the functional form of the selection model is moderately misspecified; under misspecification, the bias of estimates is inflated, and coverage degrades below the 95% level, but the method is largely still able to recover the correct population-level parameters.

Table 2. Parameter Recovery Comparison by Parameter Category: Baseline Parameter Setting

	MPL		AGG	
Parameters	MAPB (%)	IQR (%)	MAPB (%)	IQR (%)
Heterogeneity means	0.0%	13.6%	61.6%	188.0%
Heterogeneity variances	0.7%	7.7%	168.7%	694.7%
Heterogeneity correlations	8.0%	60.0%	152.0%	709.0%
Homogeneous parameters	0.6%	3.4%	26.2%	99.9%
Selection parameters	1.4%	11.2%		

Table 3. Parameter Recovery Comparison by Parameter Category: Baseline Parameter Setting

Parameters	Mean absolute bias (%)	Coefficient of variation (%)	Coverage (95% confidence interval)
Heterogeneity means	1.0%	11.0%	95.0%
Heterogeneity variances	0.6%	6.8%	95.0%
Heterogeneity correlations	6.8%	42.1%	98.3%
Homogeneous parameters	0.2%	3.2%	95.0%
Selection parameters	1.4%	10.9%	98.0%

6. Application to Spotify

Next, we apply the model to data on paying subscriber activity at Spotify (NYSE: SPOT), the world's largest music streaming platform. The vast majority of Spotify's revenues come from these subscribers, who pay a monthly fee for access to services.[17]

Spotify publicly discloses *END* and *LOSS* data (Equation (7)) in investor presentations and SEC filings. The data are left- and intermediate-censored; whereas commercial operations commenced in October 2008 (i.e., $m = 0$), Spotify began disclosing *END* data intermittently in Q1 2011, began disclosing *END* data every quarter in Q1 2015, and began disclosing *LOSS* data every quarter in Q4 2015. We model these data through Q3 2018 (i.e., the number of months in the calibration period $M = 120$).

In addition to these public disclosures, Second Measure provided us with a credit card panel data set. This data set consists of the monthly transaction activity data for 3,003,746 panel members from January 2015 (i.e., $m^* = 75$) to September 2018. A total of 289,541 of these panel members were initially acquired as Spotify premium subscribers at some point during the observation period. Spotify's publicly disclosed customer data are given in Online Appendix 7, and additional detail regarding the panel data is given in Appendix A.

We use the same three time-varying covariates in our four submodels: Quarterly dummy variables to capture seasonal fluctuations in the propensity to add and drop service throughout the year. Spotify's service is offered to individuals, and although the company expanded into new geographies in a staged manner, they have operated globally since 2011. Therefore, our unit of analysis is an individual person, and our population is the world population, as this represents everyone who could possibly acquire Spotify's service.

As in the previous section, we estimate the parameters via maximum proxy likelihood. Each stage of estimation was performed using the R programming language's nlm function, which uses a Newton-type optimization routine, letting the algorithm run until convergence. We initialize the first stage of nlm at an approximate solution obtained using DEoptim, an evolutionary algorithm, so that our starting parameter values for nlm were in a better part of the parameter space.

6.1. Model Assessment and Comparison

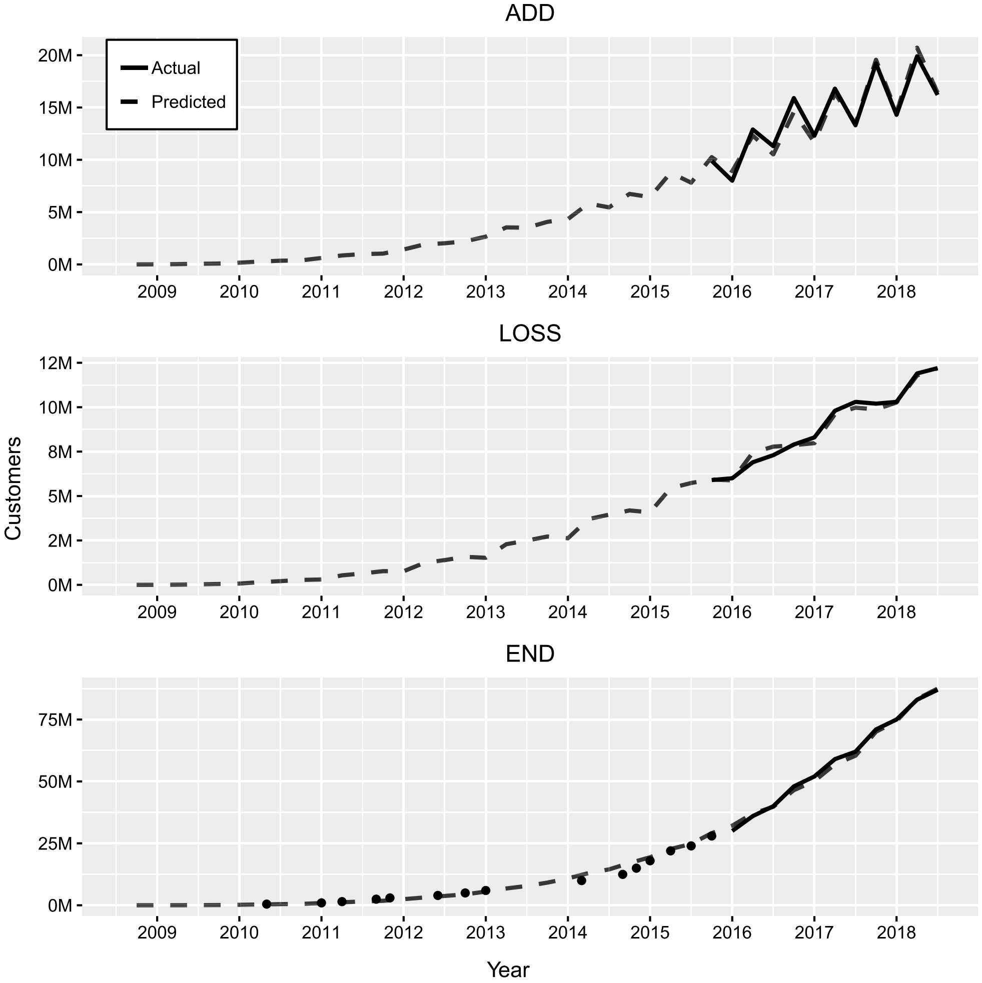
We first validate the method by examining its in- and out-of-sample performance. To evaluate in-sample performance, we fit the proposed model to all Spotify data and then plot the observed aggregate data—*ADD*, *LOSS*, and *END*—with its corresponding model-based prediction. The resulting plots are shown in Figure 3. The in-sample fit of the proposed model is good: Errors are small with no systematic pattern of under- or overprediction.

Although the in-sample fit of the proposed model is good, it tells us little about the model's predictive validity, whether credit card panel data improves predictive validity, or how the model's forecasting accuracy compares to that of extant CBCV models. Moreover, although Spotify regularly discloses *ADD*, *LOSS*, and *END*, many more companies only regularly disclose *END* (e.g., Netflix, Blue Apron, HelloFresh, and Care.com), so it is informative to assess how predictive validity varies as a function of what data are used for estimation. Predicting *ADD* and *LOSS* when only *END* is observed at the population level also provides us insight into how much the panel data improves our ability to infer measures that are not directly observed in the aggregate data—given the importance of teasing apart initial and repeat behaviors (which are never directly observed in the aggregate data), this is highly important as well. To better understand these questions, we run a rolling holdout analysis. We study the performance of the proposed model as a function of what data are available by varying the observable training data along two dimensions:

1. Aggregated data: We either train on all aggregated data, on *END* data alone, or on no aggregated data.
2. Panel data: We either train on the panel data, or we do not.

We consider the resulting five nondegenerate data availability scenarios for the proposed model. We compare these proposed model variants to the models in Gupta et al. (2004), Schulze et al. (2012), and McCarthy et al. (2017), which we refer to hereafter as GLS, SSW, and MFH, respectively. Given the severity of the seasonal fluctuations in the observable data, we enhance the GLS and SSW specifications by incorporating time-varying covariates into them. This allows

Figure 3. Spotify: Quarterly Additions, Losses, and Total Subscribers



us to incorporate the same quarterly seasonal dummy variables into all benchmark models so that no models are penalized for their inability to capture seasonality. Details of the enhanced model specifications are provided in Online Appendix 6.

For each model, we vary the length of the calibration period M, for $M = 99, 102, \ldots, 117$, corresponding to all possible calibration periods from Q4 2016 to Q2 2018. Q4 2016 is the first quarter in which *ADD* and *LOSS* data are available for four quarters, identifying the seasonal dummy variables, making it a suitable starting point for the rolling validation.

In sum, the predictive validity of GLS, SSW, MFH, and the MPL variants are based upon rolling (up to) six-quarter-ahead predictions over seven different calibration periods. For each calibration period, we predict *ADD*, *LOSS*, and *END*, resulting in 81 total rolling predictions. We summarize the predictive performance of these models by computing the MAPE of each models' predictions, averaging across all calibration periods. Table 4 provides the resulting MAPE figures.

The proposed model trained on the panel data alone (row four in Table 4) performs poorly, with MAPE figures in excess of 400% across the board. This suggests that panel selection bias is not ignorable and that naively combining the panel data with the aggregate would make the resulting forecasts worse than if the panel data were simply ignored. When *END* is observed without any panel data, the proposed model forecasts *END* reasonably well, but the corresponding MAPE figures for *ADD* and *LOSS* exceed 100%, implying that the proposed model cannot separate out the acquisition and churn processes using *END* disclosures alone.

We see uniform improvements in predictive accuracy when we add panel data, whether only *END* is observed (row 6 versus row 5) or all aggregate disclosures are observed (row 8 versus 7). Similarly, predictions uniformly improve when *ADD* and *LOSS*

Table 4. Spotify: Average Holdout MAPE for All Disclosures and Models

Model	Aggregate data	Panel data	*ADD*	*LOSS*	*END*
GLS	All	No	23.4%	31.0%	22.1%
SSW	All	No	25.3%	24.3%	6.7%
MFH	All	No	14.4%	8.9%	7.1%
Proposed	None	Yes	481%	628%	860%
	END only	No	132.8%	211.5%	10.5%
	END only	Yes	25.3%	41.9%	8.3%
	All	No	13.1%	13.3%	6.4%
	All	Yes	13.0%	9.8%	4.7%

data are observed in addition to *END*, whether panel data are observed (row 8 versus row 6) or not (row 7 versus row 5).

Our proposed model using all aggregate and panel data is the best-performing model overall. Although its MAPE is higher than MFH with respect to *LOSS*, it has the lowest MAPE across all models with respect to *ADD* and *END*. *END* is a particularly important disclosure because it is most directly tied to total revenues.

The performance of the proposed model is robust to forecasting horizon. In Figure 4, we plot the average MAPE with respect to *END* by forecasting horizon for the proposed model and for MFH. MFH has a smaller MAPE for very short forecasting horizons, but its MAPE grows quickly as the forecasting horizon lengthens, rising to approximately 20% six quarters out. MPL's forecasting error with respect to *END* is in the single digits over all forecasting horizons. These results are relevant given the importance of long-run forecasting accuracy in CBCV settings.

Having established the predictive validity of the proposed model, we next turn to insights that can be derived from the model.

6.2. Parameter Estimates and Model Insights

The parameters of the estimated model trained on all available aggregate and panel data are shown in Table 5 (associated standard errors are provided in parentheses, estimated using the asymptotic variance formula derived in Online Appendix 3). The seasonal fluctuations evident in the first two plots of Figure 3 seem to be primarily caused by relatively high propensity of subscribers to be repeat acquired in Q2 and Q4. We observe a strong positive correlation between

Figure 4. (Color online) Spotify: Average MAPE by Forecasting Horizon for *END* Disclosures (Proposed Model vs. MFH)

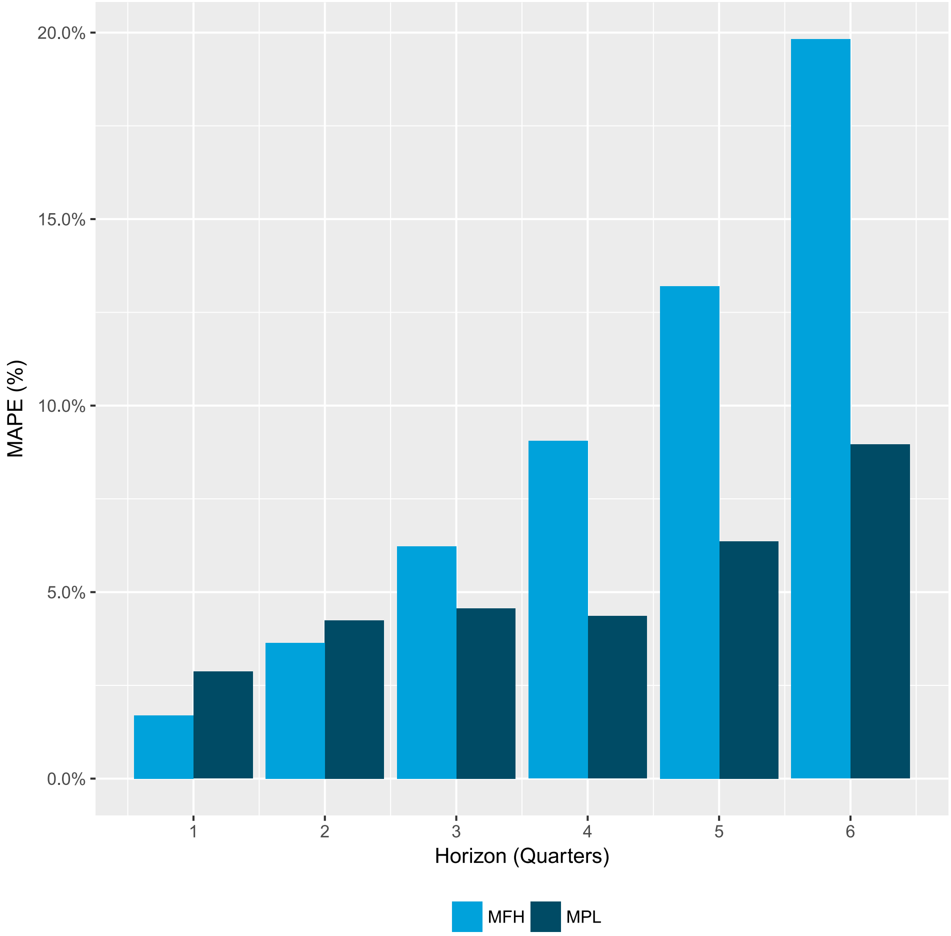


Table 5. Parameter Estimates (Standard Errors): Spotify

	Acquisition		Churn
Parameters			
Initial behavior			
λ_0	2.616×10^{-10} (0.490×10^{-10})		0.012 (0.116)
c	3.369 (0.008)		1.175 (0.595)
β_{Q1}	0.250 (0.043)		−0.483 (1.714)
β_{Q2}	0.319 (0.075)		−0.105 (0.605)
β_{Q3}	0.265 (0.066)		−0.178 (0.872)
σ_λ	2.494 (0.073)		5.517 (22.748)
π_A	0.983 (0.048)		
Repeat behavior			
λ_0	8.362×10^{-5} (14.972×10^{-5})		0.031 (0.043)
c	2.286 (0.128)		0.109 (0.111)
β_{Q1}	−1.856 (0.106)		2.684 (1.283)
β_{Q2}	−0.632 (0.277)		0.602 (0.901)
β_{Q3}	−1.793 (0.806)		1.812 (2.477)
σ_λ	1.174 (1.106)		0.019 (0.035)
π_A	0.996 (0.004)		
Panel selection			
$\beta_0^{(Z)}$	0.514 (41.834)		
$\beta_{IA}^{(Z)}$	−0.156 (2.530)	$\beta_{IC}^{(Z)}$	−0.600 (2.307)
$\beta_{RA}^{(Z)}$	2.893 (1.587)	$\beta_{RC}^{(Z)}$	−2.868 (4.918)
Correlation			
$\rho_\lambda^{(IA,IC)}$	0.516 (0.152)	$\rho_\lambda^{(IC,RA)}$	0.604 (0.250)
$\rho_\lambda^{(IA,RA)}$	0.994 (0.275)	$\rho_\lambda^{(IC,RC)}$	−0.052 (2.448)
$\rho_\lambda^{(IA,RC)}$	0.035 (4.098)	$\rho_\lambda^{(RA,RC)}$	−0.005 (5.573)

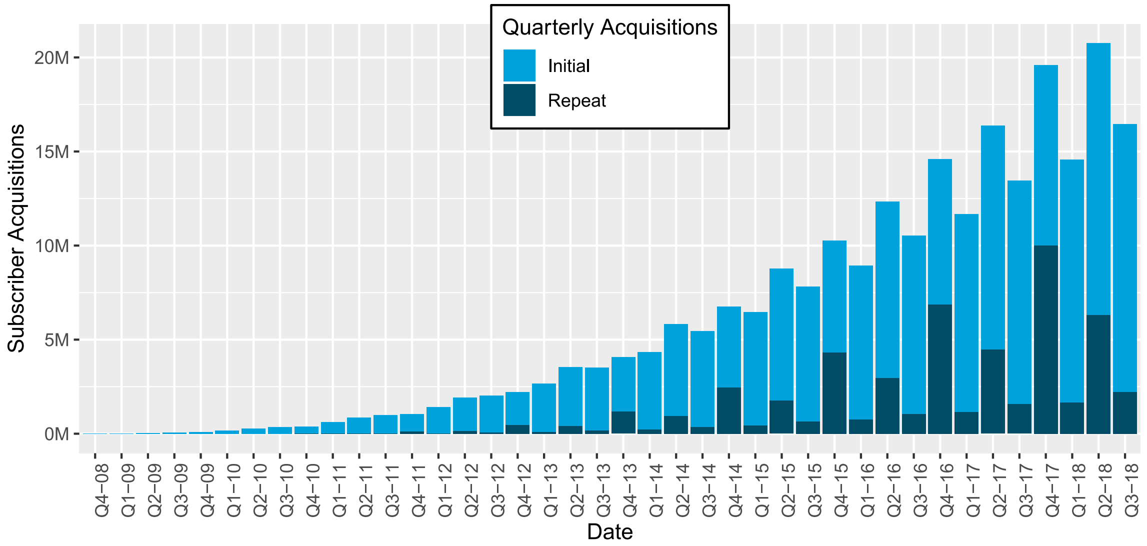
the propensity to be initially acquired and the propensity to initially churn, implying that subscribers who join later on are more likely to be loyal customers. This finding is consistent with previous work (Schweidel et al. 2008a). Customers with high propensities to be initially acquired also tend to have high propensities to be reacquired, as can be seen from the high value of $\rho_\lambda^{(IA,RA)}$. Finally, the values of $\boldsymbol{\beta}^{(Z)}$ suggest that panel members have higher propensities to readopt, and marginally lower propensities to rechurn, than the population as a whole. These selection effects may stem from the fact that the panel members are U.S.-based credit card holders and as such are wealthier and more likely to adopt than the average Spotify prospect.

The standard errors of some parameters are large, particularly those pertaining to heterogeneity and selection bias. Although some of these parameters have innocuous explanations for their standard errors,[18] this nonetheless suggests that the empirical identification of our model is not strong, even after performing data fusion. This reflects the fact that our aggregate data are limited: Although the panel is very informative about initial and repeat behaviors, under the presence of selection bias, we are uncertain as to how well this information generalizes to the population as a whole. Hence, despite Table 4 demonstrating that accounting for selection bias is essential, the limited time series of *ADD* and *LOSS* data available in our context (12 quarters) does not allow for full disentanglement of the different dimensions of selection bias. As a result, the individual selection model parameters are estimated imprecisely; in turn, the population-level estimates of the heterogeneity distribution are imprecise. These standard errors accordingly convey our uncertainty in generalizing from panel to population; conversely, ignoring selection bias would yield misleadingly precise estimates. However, the standard errors are narrower than they otherwise would be estimating on the aggregate data alone, because of the additional information gained through the inclusion of a second data source.

Turning to model insights, as we discussed in Section 1.2, repeat behaviors have significant consequences for Spotify's long-term financial health. Figure 5 plots total quarterly acquisitions, broken down between initial and repeat acquisitions. This figure shows that, although repeat acquisitions had comprised a relatively small proportion of total acquisitions historically, they have been growing significantly over time. Fully 29% of all acquisitions were from repeat acquisitions over the last 12 months of the data.

The shift in composition of Spotify's subscriber base toward reacquirers is consequential for Spotify's

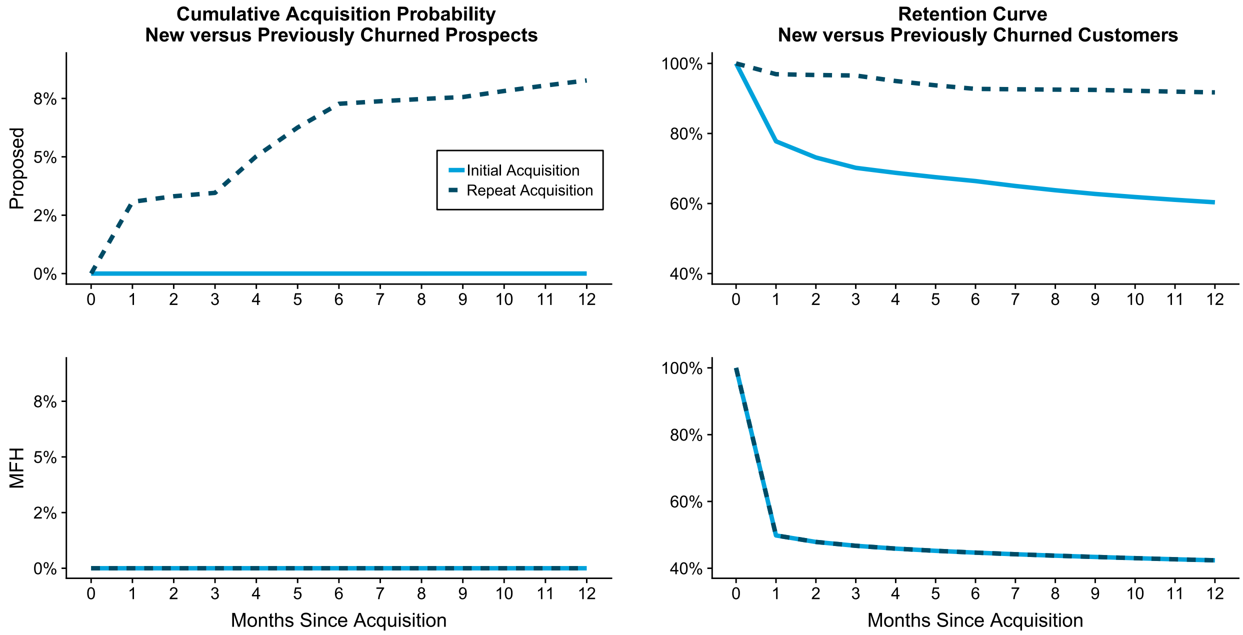
Figure 5. (Color online) Spotify: Estimated Quarterly Acquisitions by Form, Initial vs. Repeat



financial valuation for two reasons: (1) Previously churned prospects have materially higher propensities to be acquired than new prospects and (2) previously churned customers have better retention profiles than customers who have not churned yet. The upper panel of Figure 6 visualizes this, showing initial and repeat cumulative acquisition probabilities (left) and retention curves (right) as of the end of Q3 2018. In both cases, the dotted lines are well above the solid lines, indicating stronger repeat acquisition and retention than initial acquisition and retention. Although most new prospects will not be acquired within one year, approximately 8% of previously churned customers will, and although 40% of new customers will churn within one year of being acquired, the corresponding figure for previously churned customers is only 8%.

As Spotify matures, the composition of total acquisitions will continue shifting toward repeat acquisitions. This will stabilize the rate of customer acquisition

Figure 6. (Color online) Spotify: Initial and Repeat Cumulative Acquisition (Left) and Retention (Right) Curves



Notes. The upper panels correspond to the proposed model estimated with the MPL method. The lower panels correspond to the model from McCarthy et al. (2017). The first column corresponds to the cumulative acquisition probability for customers who first became a prospect in the final month of the calibration period, whereas the second column corresponds to the retention curve for customers acquired in the final month of the calibration period.

and improve Spotify's overall average retention profile. In contrast, Gupta et al. (2004) and Schulze et al. (2012) assume that zero customers will ever be reacquired, making repeat churn irrelevant. McCarthy et al. (2017) allow for reacquisition but assume that repeat acquisition and repeat churn propensities are identical to the corresponding (worse) initial acquisition and churn propensities. As a result, all three alternative models will understate total reacquisitions and the growth potential of Spotify's customer base as a whole. This is evident from just how much the repeat acquisition and retention curves implied by the proposed model (dotted lines in the upper panels) lie above the corresponding repeat acquisition and retention curves for MFH (lower panels of Figure 6). By way of example, the implied 12-month retention rate for reacquired customers is only 42.4% under MFH, well below an implied 91.7% under our proposed model.

Without any panel data to identify individual-level customer dynamics, models such as MFH are forced to make simplifying assumptions because, as we have seen through simulations and the rolling validation, aggregated data alone can only accurately model and forecast metrics that are directly historically observed. Although these assumptions are necessary for identification when using only aggregate data, as evidenced here, they can lead to substantial biases in long-term growth projections. By incorporating the panel data into our model through our proposed method, we are able to separate out initial and repeat behaviors, thus overcoming this limitation. Although our empirical estimates are still imprecise for some parameters even after data fusion, the panel data nonetheless improved our ability to make inferences and predictions compared with exclusively using aggregate data.

7. Discussion

It is increasingly common for modelers to face situations in which there is more than one data set that is available for a problem at hand. In this paper, we provide a tool for these situations, which allows modelers to use as much data as possible while doing so in a way that accounts for the differing degrees of aggregation and potential selection bias in the underlying data sources. Our proposed estimation method, maximum proxy likelihood (MPL), allows for statistically efficient estimation of models on multiple sources of data while correcting for selection bias, leading to better predictions and inferences than using single sources of data that are either highly aggregated or suffer from selection bias.

The data fusion methodology proposed here is transferable to many other problems, both in marketing and economics. Within CBCV, an important extension would be to nonsubscription firms, as in McCarthy and Fader (2018), where churn behavior is latent. Although CBCV is prediction-focused, in other settings such as discrete choice modeling, the goal may be to infer customer-level sensitivity to price or other marketing variables. For such problems, aggregate market share data could help to generalize inferences beyond the population of household scanner panel members, who may differ from the general population in their sensitivities even after controlling for demographics (Lusk and Brooks 2011). The simulations and identification analyses we performed suggest that the proposed methodology would be well-suited to such inference problems.

Although the model specifications and computations required for these other settings differ from ours, the same approximation and selection correction methods can be used. For models with Markovian structure, belief propagation algorithms such as the one we derive in Online Appendix 2 can be used to efficiently compute the moments required to use MPL.

Our proposed methodology could also be applied to other data structures. For example, it is often the case that companies only possess detailed internal transactional data for recently acquired cohorts of customers (e.g., because of adoption of a new CRM record system), but possess aggregated statistics summarizing customer activity of previous cohorts. In this case, our method can be used to estimate models for the whole customer base, and in some ways the method would be easier to apply because the selection mechanism determining Z_i is known. It could also be the case that multiple partially overlapping panel data sets are available (e.g., a combination of credit card panel data and clickstream data), and/or that nonrepresentative aggregate data are available (e.g., statistics reported by a market research firm). Our method can be further generalized to incorporate several data sources, each of which may have different selection mechanisms.

Of course, applying our method to more general settings requires careful consideration of model identification. In other data settings with different classes of models, principles similar to our identification arguments in Section 4 still hold: Our method requires disaggregate data rich enough that it helps identify the individual-level processes that are difficult to observe directly in aggregate data, and requires that there is a representative data source that is rich enough to capture population differences from the nonrepresentative data sources along relevant dimensions of the process being modeled. The complexity of the individual-level behavioral model and the selection model required will depend on the context, and the amount of data required to identify the model will vary accordingly; as discussed in Section 4.3, Our method can be used to estimate models with more complex selection mechanisms, but this in turn requires access to more extensive representative data sources that can tease apart different dimensions of selection bias.

Finally, the theoretical treatment of panel selection could be further enriched. For example, the degree to which including panel data improves performance relies in part on having nonnegligible overlap between the distribution of the individual-level rate parameters λ for the panel members and the corresponding distribution for the population members—otherwise, inferences could be based on extrapolations from panel members who are outliers relative to the broader population. An important consideration for future work is the development of benchmarks to assess whether there is sufficient overlap between the panel and target population to allow for reliable identification, analogous to methods for assessing the overlap condition in causal inference (Imbens and Rubin 2015). In the absence of a formal overlap measure, we advocate thoughtful model validation to empirically assess whether the inclusion of panel data improves performance (e.g., the rolling predictive validation we performed in the Spotify example).

Looking forward, we hope that this paper encourages analysts to more actively seek out new data sources by arming them with a framework to incorporate these varied data sources into their models. As the diversity of available data sources grows, the need for data fusion methodologies such as the one proposed in this paper will grow with it.

Acknowledgments

The authors contributed equally to this work. They thank Second Measure for providing access to their credit card panel data; Eric Bradlow and Peter Fader for discussion and guidance throughout the research process; and Shane Jensen, Kinshuk Jerath, Andrey Simonov, Oded Netzer, Olivier Toubia, Asim Ansari, Benjamin Levine, and Matteo Alleman for helpful suggestions. The authors also thank Emory University's Research Computing team for access to and support in the use of computing resources. The authors do not have any financial interest, direct or indirect, in the company studied in the manuscript.

Appendix A. Spotify Implementation Details

In this appendix, we provide additional details on our application using Spotify and Second Measure data.

1. We assume annual formation of new prospect pools and compute the size of each prospect pool based on the world population at and after the time of Spotify's incorporation. That is, we define the size of the initial prospect pool (born at Spotify's time of incorporation in 2008) as the global population in 2008, the prospect pool born in 2009 as the net growth in global population from 2008 to 2009, and so on. We assume that all panel members in the Second Measure data are drawn from the initial 2008 prospect pool.

2. World population data come from the World Bank (https://data.worldbank.org/indicator/SP.POP.TOTL). This data set contains the world population each year through 2017, whereas the aggregate and disaggregate Spotify data runs through and including Q3 2018, and projections of future customer behavior requires projection of the world population further into the future. As in McCarthy et al. (2017) and McCarthy and Fader (2018), we run a time series regression using world population data (from 1963 to 2017) to forecast the world population in future years. The world population was strictly increasing over this period. Although a simple linear regression of year-on-year percentage change in world population by year does not reject the null hypothesis of the augmented Dickey-Fuller test of nonstationarity, fitting the data to an ARIMA(0,1,0) model via maximum likelihood rejects the null hypothesis that a unit root is present (test statistic: −4.41, $p < 0.01$). Therefore, we use an ARIMA(0,1,0) specification, which has an $R^2 = 99.0\%$.

3. Spotify provides a promotional offer to new customers in the second and fourth quarters of each calendar year, allowing prospects to pay a discounted price upfront to trial the service for the next three months. Virtually all trial amounts are less than $1.30, far below Spotify's regular price of $9.99 per month. As such, there are a number of panel members with a spend amount of $1.30 or less in one month, followed by no payments in the next two months (while the trial offer is still in effect). We assume that subscribers are retained during the promotional period.

4. The data we obtained from Second Measure do not include any panel members who churned from the panel during the observation period. The proposed approach could be extended to data sets with panel attrition by incorporating an additional timing process for the duration between when panel members enter the panel to when they leave the panel, as long as the date of panel entry was also observed.

5. There were no new panel members acquired into the data set during the observation period.

6. The credit card panel data set has a number of so-called *skips*, or months for which there is no payment, despite there being payments in the month immediately before and after. Second Measure noted that these skips are often because of the timing of when payments are processed versus when they are charged. For this reason, we assume that customers are retained in single-month skips. Accordingly, our model incorporates a one-month lag between when a customer churns and when they are first eligible to be reacquired: In particular, if a customer churns in month m (churn defined as the first month in which there is not a payment), they are *reborn* as a prospect in month $m + 1$, and are first eligible to be reacquired in month $m + 2$. As such, our model specification is in concordance with the assumption that customers are retained in single-month skips: In our specification, a churn entails at least two months of dormancy.

7. Our panel data set is left-truncated. Second Measure's panel dates back to January 2011, but they provided us with granular data that begins in January 2015 for just the panel members who made no purchases at Spotify from when Second Measure's data began in January 2011 through December 2014 and then registered their first payment in the data set during the observation period. We assume that all active panel members were initially acquired in the first month we observe a payment

at Spotify in the panel data set.[19] The probability that these customers made a purchase at Spotify before January 2011 is minimal because Spotify acquired very few customers this early on.

8. Second Measure also provided us with the total size of their panel, whether or not those panel members made purchases during the observation period. As described in Online Appendix 2, we account for the fact that panel members who do not make purchases during the observation period were either inactive before and during the observation period (i.e., made no purchases at all prior to October 2018), or only were acquired before the observation period began (i.e., made their first purchase before January 2015, and thus were excluded from our granular panel data set).

Endnotes

[1] Furthermore, our method retains favorable statistical properties even when only the first moment is computable.

[2] As such, our method does not apply to macro models, such as time series models for aggregate data. For macro models, researchers should consider other data fusion methods.

[3] We could add individual-specific covariates $\mathbf{x}_i$ into this specification as well. Doing so would require knowing the population-level distribution of $\mathbf{x}_i$.

[4] This specification allows for substantial time dynamics in customer acquisition and retention profiles, because of correlations between dimensions of $\boldsymbol{\lambda}_i$ and separate specification of initial and repeat processes. It could be tempting to further enrich the specification of $\boldsymbol{\lambda}_i$ (e.g., allowing parameters to evolve over time); however, these extensions are likely to be confounded with existing sources of dynamics, which could hurt performance due to poor identification.

[5] To do so, we would need to know the population distribution of the covariates, because computing the aggregate moments requires us to integrate out the population distribution of covariates. Under the assumption that $\boldsymbol{\xi}_i \perp\!\!\!\perp \mathbf{Y}_i|\boldsymbol{\lambda}_i$, including other covariates is unnecessary, because any selection bias will be captured by dependency with $\boldsymbol{\lambda}_i$; however, including observed covariates could improve the statistical precision with which the selection function is estimated.

[6] Although each individual in the population is allowed to have a unique parameter vector $\boldsymbol{\lambda}_i$, we marginalize the individual-level parameters out in estimation.

[7] Under stricter conditions than the central limit theorem, local limit laws further state that the likelihood of the N-fold convolution converges uniformly to the multivariate normal density (Bhattacharya and Rao 1986), such that this approximation is asymptotically exact. However, just the regularity conditions of the central limit theorem are sufficient for our estimator to achieve consistency and asymptotic normality.

[8] In particular, the covariance matrix $\boldsymbol{\Sigma}_N$ has dimension $q \times q$, where q is the dimension of $\mathbf{D}_N$; thus, the number of covariance elements that need to be computed scales quadratically in the dimension of $\mathbf{D}_N$.

[9] This is true for any summary statistics that are affine transformations of $\mathbf{Y}_i$. Although most commonly-disclosed summary statistics in SEC filings are affine transformations of $\mathbf{Y}_i$, we could also generalize to other types of summary statistics by computing moments using the delta method or simulation-based approaches (Gourieroux et al. 1993).

[10] One could also iterate between estimating $\boldsymbol{\theta}$ and updating $\hat{\boldsymbol{\Sigma}}_N$ multiple times, but this is asymptotically equivalent to the two-stage procedure and so will have the same theoretical properties. In our parameter recovery simulation study (Section 5.2), we continued estimation for a third stage and found that the mean absolute estimation error did not improve within three significant figures, versus a 7.2% improvement moving from the first stage to the second.

[11] We can also use the computed asymptotic covariance matrix to construct a prediction interval for our forecasts by iteratively sampling parameter vectors from the asymptotic distribution of the parameter estimates and then sampling realizations of the data from those sampled parameter vectors.

[12] Extensive simulations supporting these arguments are available upon request.

[13] For instance, if the aggregate data contain spikes in $LOSS_q$ after spikes in ADD_q in early periods, whereas the panel data do not show as pronounced spikes, this suggests that there is a long tail of $\lambda_i^{(IC)}$s in the aggregate data but that the panel distribution has lighter tails, suggesting that people with high $\lambda_i^{(IC)}$ are underrepresented in the panel.

[14] Indeed, in our simulations reported in Section 5.2, we find that when using aggregate data alone, the means of the heterogeneity distributions are much more precisely estimated than the variances. In the absence of second-order selection biases, the panel data can aid in the identification of the population heterogeneity variances.

[15] Although not part of the formal simulation study, we see the same pattern of relative performance across methods when we assume that only *END* data are observed historically and are forecasting *END* versus *ADD* and *LOSS*.

[16] The large-scale simulation study in the previous section had explicit baseline data setting levels (e.g., $M = 60$ and $N = 100$K), so we left those settings as-is for this exercise. The simulation had low and high values for each parameter, so our baseline scenario for this exercise averages these two values for each parameter.

[17] On a trailing 12-month basis, approximately 90% of Spotify's total revenues was generated from premium subscriber fees in Spotify's eight most recent quarters. This proportion has remained relatively constant over time.

[18] The standard error associated with the $\beta_0^{(Z)}$ coefficient is large relative to its point estimate because of uncertainty in the mean of the distribution of $\boldsymbol{\lambda}_i$. If we were to center the $\boldsymbol{\lambda}_i$s in the selection equation, the point estimate for $\beta_0^{(Z)}$ in the centered equation would be −10.60 with a standard error of 2.58. Additionally, the pairwise correlation parameters between $\lambda^{(RC)}$ and the other $\lambda^{(p)}$ terms have high standard errors because $\sigma_\lambda^{(RC)}$ is small; there is little variation in $\lambda^{(RC)}$ to identify correlations with the other $\lambda^{(p)}$ terms.

[19] This is analogous to the approach used to address the initial condition problem by Erdem and Keane (1996), who use the first two years of their data set to approximate the past purchase history of panel members.

References

Bayer E, Tuli KR, Skiera B (2017) Do disclosures of customer metrics lower investors' and analysts' uncertainty but hurt firm performance? *J. Marketing Res.* 54(2):239–259.

Berry S, Levinsohn J, Pakes A (2004) Differentiated products demand systems from a combination of micro and macro data: The new car market. *J. Political Econom.* 112(1):68–105.

Bhat C (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Res. Part B: Methodological* 35(7):677–693.

Bhattacharya RN, Rao RR (1986) *Normal Approximation and Asymptotic Expansions*, vol. 64 (SIAM, Philadelphia).

Bonacchi M, Kolev K, Lev B (2015) Customer franchise—A hidden, yet crucial, asset. *Contempory Accounting Res.* 32(3):1024–1049.

Chen Y, Yang S (2007) Estimating disaggregate models using aggregate data through augmentation of individual choice. *J. Marketing Res.* 44(4):613–621.

Dias FF, Lavieri PS, Kim T, Bhat CR, Pendyala RM (2019) Fusing multiple sources of data to understand ride-hailing use. *Transporation Res. Record* 2673(6):214–224.

Erdem T, Keane MP (1996) Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Sci.* 15(1):1–20.

Feit EM, Wang P, Bradlow ET, Fader PS (2013) Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *J. Marketing Res.* 50(3):348–364.

Gourieroux C, Monfort A, Renault E (1993) Indirect inference. *J. Appl. Econometrics* 8(suppl 1):S85–S118.

Gourio F, Rudanko L (2014) Customer capital. *Rev. Econom. Stud.* 81(3):1102–1136.

Gupta S, Lehmann DR, Stuart JA (2004) Valuing customers. *J. Marketing Res.* 41(1):7–18.

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica*. 50(4):1029–1054.

Heckman JJ (1991) Identifying the hand of past: Distinguishing state dependence from heterogeneity. *Amer. Econom. Rev.* 81(2):75–79.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK).

Little RJ, Rubin DB (2019) *Statistical Analysis with Missing Data*, vol. 793 (John Wiley & Sons, Hoboken, NJ).

Lusk JL, Brooks K (2011) Who participates in household scanning panels? *Am. J. Agricultural Econom.* 93(1):226–240.

Manchanda P, Rossi PE, Chintagunta PK (2004) Response modeling with nonrandom marketing-mix variables. *J. Marketing Res.* 41(4):467–478.

McCarthy D, Fader P (2018) Customer-based corporate valuation for publicly traded noncontractual firms. *J. Marketing Res.* 55(5):617–635.

McCarthy D, Fader P, Hardie B (2017) Valuing subscription-based businesses using publicly disclosed customer data. *J. Marketing* 81(1):17–35.

Musalem A, Bradlow ET, Raju JS (2008) Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *J. Marketing Res.* 45(6):715–730.

Musalem A, Bradlow ET, Raju JS (2009) Bayesian estimation of random-coefficients choice models using aggregate data. *J. Appl. Econometrics* 24(3):490–516.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Francisco).

Schulze C, Skiera B, Wiesel T (2012) Linking customer and financial metrics to shareholder value: The leverage effect in customer-based valuation. *J. Marketing* 76(2):17–32.

Schweidel D, Fader P, Bradlow E (2008a) A bivariate timing model of customer acquisition and retention. *Marketing Sci.* 27(5): 829–843.

Schweidel D, Fader P, Bradlow E (2008b) Understanding service retention within and across cohorts using limited information. *J. Marketing* 72(1):82–94.

Schweidel DA, Knox G (2013) Incorporating direct marketing activity into latent attrition models. *Marketing Sci.* 32(3):471–487.

Schweidel DA, Moe WW (2014) Listening in on social media: A joint model of sentiment and venue format choice. *J. Marketing Res.* 51(4):387–402.

Train K (2009) *Discrete Choice Methods with Simulation* (Cambridge University Press, Cambridge, UK).

Van Diepen M, Donkers B, Franses PH (2009) Dynamic and competitive effects of direct mailings: A charitable giving application. *J. Marketing Res.* 46(1):120–133.