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

Daniel Minh McCarthy
Emory University, daniel.mccarthy@emory.edu

Elliot Shin Oblander
Columbia University, EOblander23@gsb.columbia.edu

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. Utilizing 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 which 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 which do not use multiple data sources.

Key words: 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 thousand 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. While 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 datasets, 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 due to the difficulty of incorporating differing granularities of data and selection bias. We propose a methodology to solve these challenges, allowing for the fusing together of (1) aggregated data about the population as a whole with (2) granular data for a possibly non-representative 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 non-representative 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 non-representative granular data – is an increasingly common one in marketing, economics, and finance. As mentioned earlier, 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 (e.g., 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, since 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 non-representative 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 US),[2] for which researchers have at their disposal a combination of (1) aggregate "macro" data that is representative of the entire population

[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.

of interest, and (2) disaggregate "micro" data that covers a possibly non-representative 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 to 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 to estimation using only non-representative disaggregate data (e.g., consumer choice models estimated upon 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.

While 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 trade-off 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 is 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 is at least somewhat representative of the population of interest. Note that while 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 the resulting standard errors can still be 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 upon 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), not internal ones (e.g., internal CRM 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 of the aforementioned papers only utilize 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. While it is intuitively appealing that 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 towards 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 significantly 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, due to limited data.

As discussed in Section 1.1, these gaps are precisely those which 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.

While Second Measure's data set is large and granular, it is a non-representative 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, 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, then close with a discussion in Section 7. We include proofs of asymptotic properties, other derivations, and data in the Web 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 re-subscribe, 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, then discuss how we account for unobserved heterogeneity in the next section. For notational simplicity, 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:

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

where

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

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: e.g., 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 $\pi^{(RA)}$.

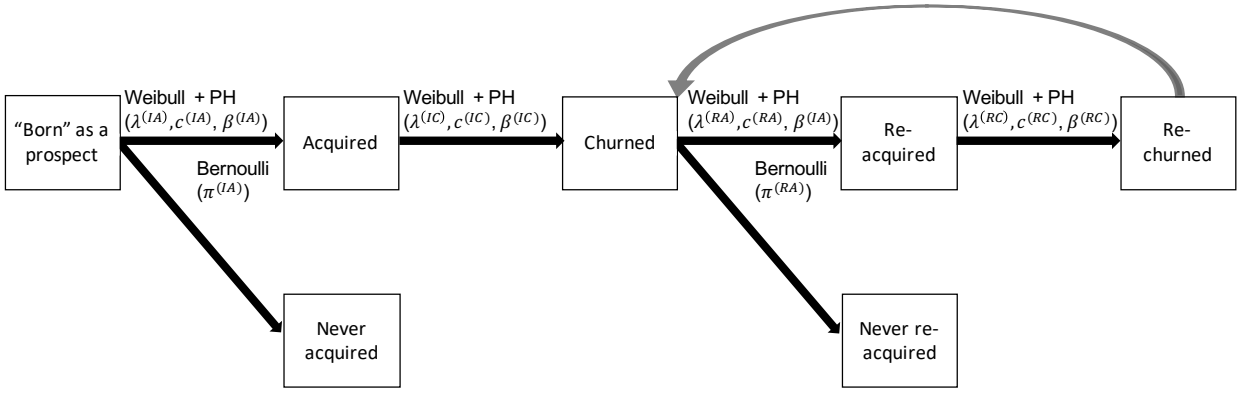
As discussed in Section 1, prior customer base analysis literature based on aggregated data has not separately modeled initial and repeat customer behavior. Incorporating granular panel data through data fusion allows us to separate out the two processes to understand how the behavioral patterns of repeat customers differ from those of first-timers. The full individual-level model is summarized 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. 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 $\boldsymbol{\lambda}_i=\left(\lambda_i^{(IA)},\lambda_i^{(IC)},\lambda_i^{(RA)},\lambda_i^{(RC)}\right)'$. While we could use independent gamma distributions for

[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 $\mathbf{x}_i$.

Figure 1 Flow diagram of the proposed individual-level acquisition and retention process



Note: 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 difference of CDF's. The Bernoulli processes determine whether or not a customer is ever (re-)acquired, while the respective Weibull processes determine time until (re-)acquisition, given that (re-)acquisition occurs.

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 which 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]

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

2.3. Panel Selection

There may be selection bias which 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 non-ignorable 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.

[4] This specification allows for substantial time dynamics in customer acquisition and retention profiles, due to 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.

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 $\log(\boldsymbol{\lambda}_i)$:

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

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 non-ignorable 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 employed by Manchanda et al. (2004), Van Diepen et al. (2009), and Schweidel and Knox (2013), who correct for non-random targeting of direct marketing by modeling targeting as a function

[5] To do so, we would need to know the population distribution of the covariates, since 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, since 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.

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 Web 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:

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

[6] While 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.

As such, the length of the observation period in the panel data is $M - m^*$ 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 $\mathbf{Y}_i$:

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

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:

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

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 on 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 (e.g. 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, while getting around scalability issues, results in poor statistical efficiency, since subsampling the data inflates estimation variance. For instance, scaling down the aggregate data by 1,000 times would result in standard errors that are $\sqrt{1000} \approx 31.6$ 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, due to the above 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:

$$\boldsymbol{\mu}_N(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\mathbf{D}_N | z_{1:N}, \tilde{\mathbf{y}}_{1:N}] \qquad \boldsymbol{\Sigma}_N(\boldsymbol{\theta}) = Var_{\boldsymbol{\theta}}[\mathbf{D}_N | z_{1:N}, \tilde{\mathbf{y}}_{1:N}] \tag{9}$$

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). Note that a typical moment-based estimator would require only the unconditional moments of $\mathbf{D}_N$, whereas here we condition

[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.

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, since 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, as 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):

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

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, then discuss our proposed estimation procedure in Section 3.4.

3.3. Computing the Proxy Likelihood

While 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

[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 $\mathbf{D}_N$.

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]

Since $\mathbf{Y}_i$ 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 very short time horizons. To make the marginalization feasible, we construct a recursive belief propagation algorithm which exploits the Markovian structure of our model, drastically 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 non-conjugate; 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 on $\boldsymbol{\lambda}^{(k)}$, then averaging over the K draws before plugging the results into the proxy likelihood equation. Since the second and third terms of Equation 10 involve probabilities and expectations conditional on the panel selection outcome Z_i and the panel data $\tilde{\mathbf{Y}}_i$, for these terms we integrate over the posterior $g(\boldsymbol{\lambda}_i|Z_i)$ and $g(\boldsymbol{\lambda}_i|Z_i, \tilde{\mathbf{Y}}_i)$ respectively 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 Web 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.

[9] This is true for any summary statistics that are affine transformations of $\mathbf{Y}_i$. While virtually all commonly disclosed summary statistics in SEC filings are affine transformations of $\mathbf{Y}_i$, our method could also generalize to non-affine transformations by using simulation-based approaches, albeit at greater computational cost (Gourieroux et al. 1993).

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 4-dimensional Halton sequence, simulate $\boldsymbol{\lambda}^{(k)}$ from the mixing distribution

$$\log\left(\boldsymbol{\lambda}^{(k)}\right) \sim \mathcal{N}\left(\log\left(\boldsymbol{\lambda}_0\right), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\right)$$

Compute the probability of selection into the panel given $\boldsymbol{\lambda}^{(k)}$ (Equation 4):

$$p_k \coloneqq P_\theta(Z = 1|\boldsymbol{\lambda}^{(k)}) = Logit^{-1}\left(\beta_0^{(Z)} + \left(\boldsymbol{\beta}^{(Z)}\right)' \log\left(\boldsymbol{\lambda}^{(k)}\right)\right)$$

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

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

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 Web Appendix 2.3:

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

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 Web Appendix, then plugging this into the multivariate normal log-density formula for aggregate data $\mathbf{d}$, yielding (up to additive constants):

$$\ell_d \coloneqq -\frac{1}{2}\left(\mathbf{d} - \boldsymbol{\mu}_N\left(\boldsymbol{\theta}\right)\right)' \left(\hat{\boldsymbol{\Sigma}}_N\right)^{-1} \left(\mathbf{d} - \boldsymbol{\mu}_N\left(\boldsymbol{\theta}\right)\right)$$

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

$$\tilde{\ell} \coloneqq \ell_z + \ell_y + \ell_d$$

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:

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

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

$\hat{\boldsymbol{\theta}}_N^{(2)}\left(z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}|\hat{\boldsymbol{\Sigma}}_N\right)$ 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 $\hat{\boldsymbol{\Sigma}}_N$:

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

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

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

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

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

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 $O_p\left(N^{-1/2}\right)$ and achieving asymptotic normality (derivations are available in Web Appendix 3).[10] We will refer to these estimation methods collectively as maximum proxy likelihood (MPL). While 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 Web 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: 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.

[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 variance matrix to construct a prediction interval for our forecasts by iteratively sampling parameter vectors from the asymptotic distribution of the parameter estimates, then sampling realizations of the data from those sampled parameter vectors.

4. Identification

We demonstrate 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, 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). While our aggregate data does not distinguish between first-time and repeat activity, it 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 due to 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 which are difficult to verify, and some degree of misspecification is inevitable in practice. As such, while 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 re-acquired. 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, since 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.

While 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), while high variance suggests a steep initial dropoff followed by a quick flattening out due to 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. due to 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. Note that the following arguments imply identification of the heterogeneity distribution of the *panel members*, 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 $\mathbf{\Sigma}_\lambda$) 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 is most informative.

While 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, while 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 above that the population mean of $\boldsymbol{\lambda}_i$ is readily identified in the aggregate data, while 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 to 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. Lastly, 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

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

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 Web Appendix 5.4, where the true data generating process has a nonlinear selection function while 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.

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

Formally, we show in Web Appendix 4 that as long as the panel data is sufficiently rich, a general selection mechanism $P_{\boldsymbol{\psi}}(Z_i = 1|\boldsymbol{\lambda}_i)$, parameterized by vector $\boldsymbol{\psi}$, is identified so 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 is 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 trade-off 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, since the moments needed to identify them will be relatively noisy compared to those needed to identify first-order selection bias.[14] Unless we have strong reason to believe that misspecification due to second-order selection biases is severe, a linear selection model is likely to perform better, as it can capture first-order shifts between the panel and aggregate data, but can still be relatively precisely estimated. In our empirical

[13] For instance, if the aggregate data contains spikes in $LOSS_q$ after spikes in ADD_q in early periods, while the panel data does 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.

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. While this is in no way a guarantee that our model does not suffer from 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 to 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 so has small incremental cost of implementation in terms of computation. Compute times by method are reported in Web 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 $M = 60$, $N = 100$K, the 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, [5-6] initial and repeat process variance, and [7-8] 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 Web 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 is observed each quarter (which fully determine ADD_q, since $ADD_q = END_q - END_{q-1} + LOSS_q$). In this section and the next, we evaluate the methods on the basis of predictive accuracy and parameter recovery, respectively, to

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%

Disclosure	PAN	AGG	MPL
ADD	17.8%	2.4%	2.0%
$LOSS$	14.5%	2.6%	2.1%
END	84.6%	0.9%	0.5%

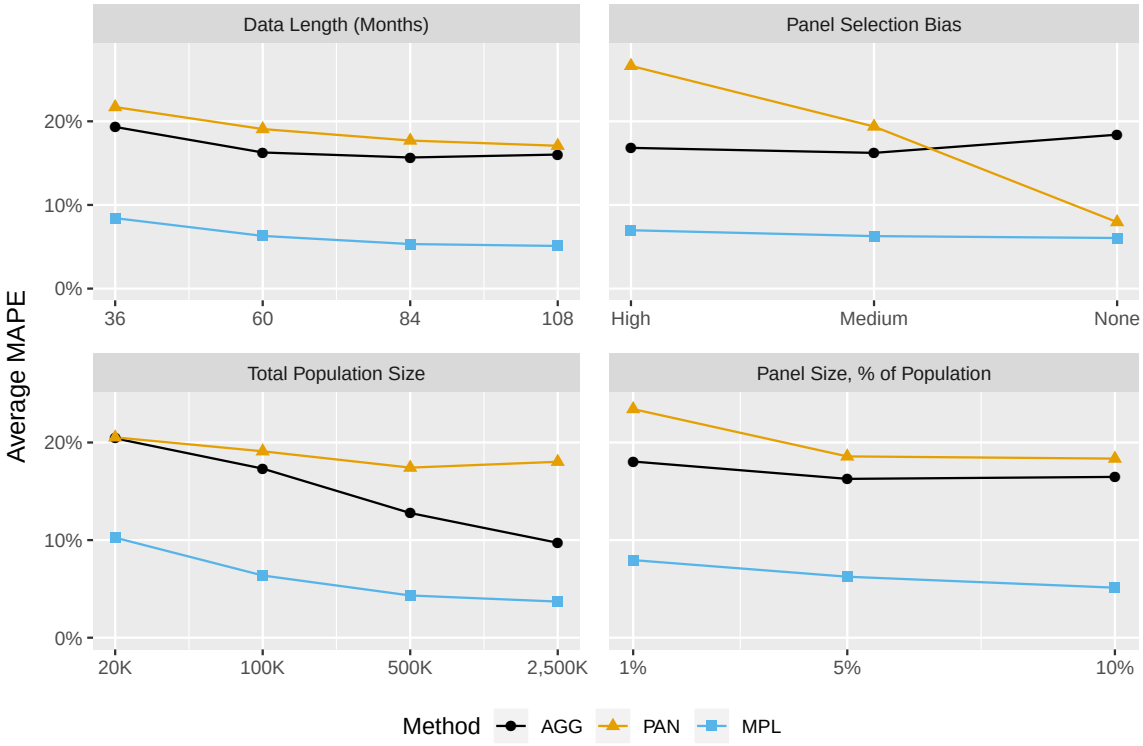
establish the usefulness of the method for both prediction-oriented use cases (e.g., CBCV) and inference-oriented ones.

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 separate out initial and repeat behavior, while 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 significant 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 the performance for QIA, QIC, QRA, and QRC. Disclosure-specific MAPE data is available in Web Appendix 5.3. MPL

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

Figure 2 **Simulation study: Holdout MAPE for QIA, QIC, QRA, and QRC by method and data setting, averaging across parameter settings and disclosures**



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.

5.2. Parameter Recovery

While 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 Web Appendix 5.4).

[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.

Table 2 Parameter recovery comparison by parameter category: baseline parameter setting

Parameters	MPL		AGG	
	Med. Abs. Bias (%)	IQR (%)	Med. Abs. Bias (%)	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%		

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 5 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 ($\boldsymbol{\beta}^{(Z)}$) – and report each performance measure averaged across all parameters within each category (parameter-by-parameter results are available in Web 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:

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

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. While 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.

Table 3 Parameter recovery comparison by parameter category: baseline parameter setting

Parameters	Mean Abs. Bias (%)	Coeff. of Variation(%)	Coverage (95% CI)
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%

Next, we focus our study upon 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 Web 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 Web 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.

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 is left- and intermediate-censored – while 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

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

disclosing $LOSS$ data every quarter in Q4 2015. We model this 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. 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 is given in Web 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 sub-models: 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 while 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.

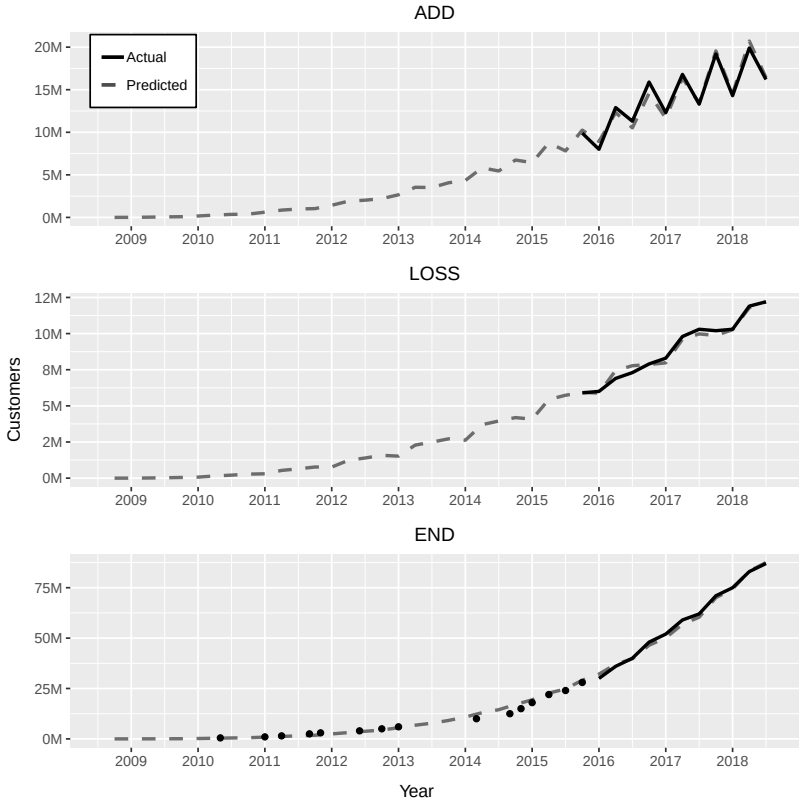
As in the previous section, we estimate the parameters via maximum proxy likelihood. Each stage of estimation was performed using the `R` programming language's `nlm` function, which uses a Newton-type optimization routine, letting the algorithm run until convergence. We initialize the first stage of `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, 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 over-prediction.

While 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, while 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 is used for estimation.

Figure 3 Spotify: Quarterly additions, losses, and total subscribers



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 is 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 non-degenerate possible 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 by 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

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%

covariates into them. This allows 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, making the resulting performance metrics more comparable. Details of the enhanced model specifications are provided in Web 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 is 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 on 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 mean absolute percentage error (MAPE) of each models' predictions, averaging across all calibration periods. Table 4 provides the resulting MAPE figures.

The proposed model trained upon 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 $LOSS$ data is observed in addition to END, whether panel data is 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. While 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.

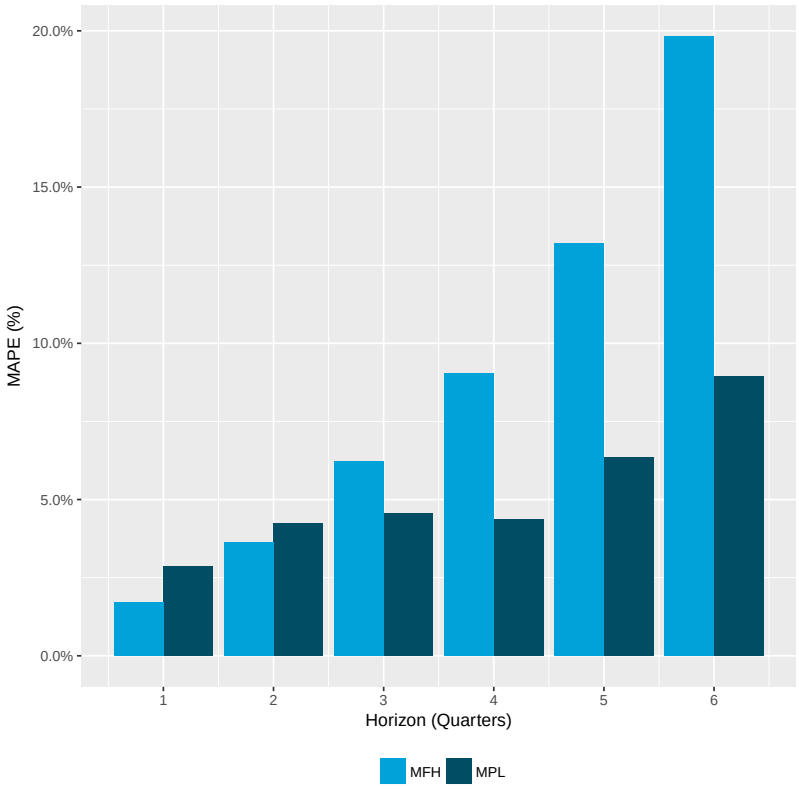
The performance of the proposed model is robust to forecasting horizon. In Figure 4, we plot the average MAPE with respect to END by forecasting horizon for the proposed model and for MFH. MFH has a smaller MAPE for very short forecasting horizons, but its MAPE grows quickly as the forecasting horizon 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 upon 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 Web Appendix 3). The seasonal fluctuations evident in the first two plots of Figure 3 seem to be primarily due to relatively high propensity of subscribers to be repeat acquired in Q2 and Q4. We observe a strong positive correlation between 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 re-adopt, and marginally lower propensities to

Figure 4 **Spotify: Average MAPE by forecasting horizon for END disclosures (proposed model versus MFH)**



re-churn, 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.

We note that the standard errors of some parameters are large, particularly those pertaining to heterogeneity and selection bias. Though 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 is limited: even though 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

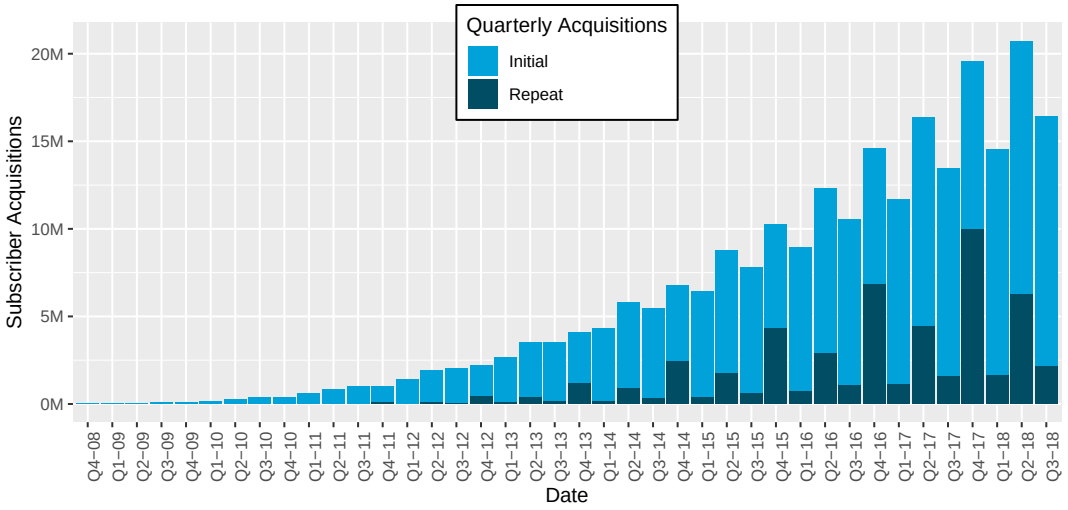
[18] The standard error associated with the $\beta_0^{(Z)}$ coefficient is large relative to its point estimate because of uncertainty in the mean of the distribution of $\boldsymbol{\lambda}_i$. If we were to center the $\boldsymbol{\lambda}_i$s in the selection equation, the point estimate for $\beta_0^{(Z)}$ 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.

Table 5 Parameter Estimates (Standard Error): Spotify

Initial Behavior Parameters			
	Acquisition		Churn
λ_0	2.616 $\times 10^{-10}$ (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 Parameters			
	Acquisition		Churn
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 Parameters			
$\beta_0^{(Z)}$	0.514 (41.834)	$\beta_{IA}^{(Z)}$	-0.156 (2.53)
$\beta_{IC}^{(Z)}$	-0.600 (2.307)	$\beta_{RA}^{(Z)}$	2.893 (1.587)
$\beta_{RC}^{(Z)}$	-2.868 (4.918)		
Correlation Parameters			
$\rho_\lambda^{(IA,IC)}$	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)

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

Figure 5 **Spotify: Estimated quarterly acquisitions by form, initial versus repeat**



uncertainty in generalizing from panel to population; conversely, ignoring selection bias would yield misleadingly precise estimates. Still, 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 while 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 past 12 months.

The shift in composition of Spotify's subscriber base towards reacquirers is consequential for Spotify's 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. While most new prospects will not be acquired within one year, approximately 8% of previously churned customers will, and while 40% of new customers will churn within one year of being acquired, the corresponding figure for previously churned customers is only 8%.

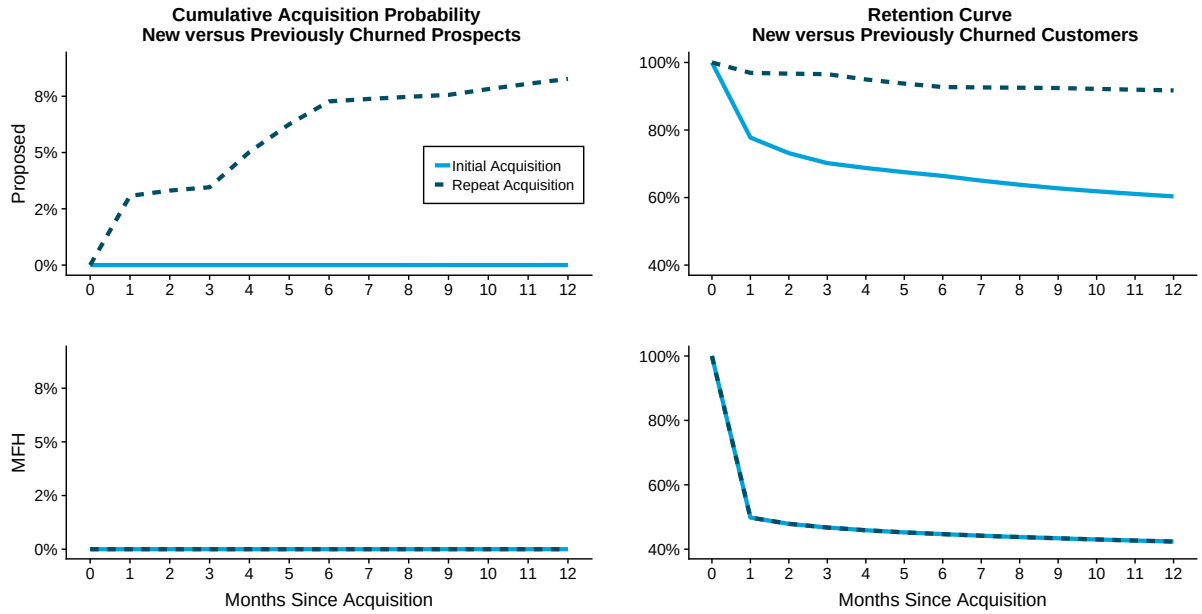
As Spotify matures, the composition of total acquisitions will continue shifting towards repeat acquisitions. This will stabilize the rate of customer acquisition and improve Spotify's overall average retention profile. In contrast, Gupta et al. (2004) and Schulze et al. (2012) assume that zero customers will ever be reacquired, making repeat churn irrelevant. McCarthy et al. (2017) allows for reacquisition, but assumes 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 blue lines in the upper panel) lie above the corresponding repeat acquisition and retention curves for MFH (lower panel 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 which are directly historically observed. While 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. While our empirical estimates are still somewhat imprecise even after data fusion, the panel data nonetheless improved our ability to make inferences and predictions compared to 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.

Figure 6 Spotify: Initial and repeat cumulative acquisition (left) and retention (right) curves



Note: The upper panel corresponds to the proposed model estimated with the MPL method. The lower panel corresponds 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, while the second column corresponds to the retention curve for customers acquired in the final month of the calibration period.

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 non-subscription firms, as in McCarthy and Fader (2018), where churn behavior is latent. While CBCV is prediction-focused, in other settings such as discrete choice modeling, the goal may be to infer customer-level sensitivity to price and 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.

While 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 Web 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. due to adoption of a new CRM record system), but possess only 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 non-representative aggregate data is 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 non-representative data sources along relevant dimensions of the process. 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 non-negligible overlap between the distribution of the individual-level rate parameters $\boldsymbol{\lambda}$ of the panel members and the corresponding distribution for the population members – otherwise, inferences could be based upon 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 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 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. We thank Second Measure for providing access to their credit card panel data. We acknowledge financial support provided by a research grant from the Goizueta Business School of Emory University. We are deeply grateful to Eric Bradlow and Peter Fader for their discussion and guidance throughout the research process. We are also thankful to Shane Jensen, Kinshuk Jerath, Andrey Simonov, Oded Netzer, Olivier Toubia, Asim Ansari, Benjamin Levine, and Matteo Alleman for helpful suggestions.

Web 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 comes from the World Bank (`https://data.worldbank.org/indicator/SP.POP.TOTL`). This data set contains the world population each year through 2017, while 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. While 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 non-stationarity, 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-value $< .01$). 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 every Q2 and Q4, allowing prospects to pay a discounted price up-front 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 does 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 due to 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 re-acquired: in particular, if a customer churns in month m (churn defined as the first month in which there is not a payment), they are "re-born" as a prospect in month $m+1$, and are first eligible to be re-acquired 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, 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 prior to 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 Web 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).

References

Bayer E, Tuli KR, Skiera B (2017) Do disclosures of customer metrics lower investors' and analysts' uncertainty but hurt firm performance? *Journal of Marketing Research* 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. *Journal of Political Economy* 112(1):68–105.

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

[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 dataset to approximate the past purchase history of panel members.

Bhattacharya RN, Rao RR (1986) *Normal approximation and asymptotic expansions*, volume 64 (SIAM).

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

Chen Y, Yang S (2007) Estimating disaggregate models using aggregate data through augmentation of individual choice. *Journal of Marketing Research* 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. *Transportation Research 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 Science* 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. *Journal of Marketing Research* 50(3):348–364.

Gourieroux C, Monfort A, Renault E (1993) Indirect inference. *Journal of Applied Econometrics* 8(S1):S85–S118.

Gourio F, Rudanko L (2014) Customer capital. *Review of Economic Studies* 81(3):1102–1136.

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

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.

Heckman JJ (1991) Identifying the hand of past: Distinguishing state dependence from heterogeneity. *The American Economic Review* 81(2):75–79.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Little RJ, Rubin DB (2019) *Statistical analysis with missing data*, volume 793 (John Wiley & Sons).

Lusk JL, Brooks K (2011) Who participates in household scanning panels? *American Journal of Agricultural Economics* 93(1):226–240.

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

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

McCarthy D, Fader P, Hardie B (2017) Valuing subscription-based businesses using publicly disclosed customer data. *Journal of 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. *Journal of Marketing Research* 45(6):715–730.

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

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

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

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

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

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

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

Train K (2009) *Discrete choice methods with simulation* (Cambridge university press).

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

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

Daniel Minh McCarthy
Emory University, daniel.mccarthy@emory.edu
Elliot Shin Oblander
Columbia University, EOblander23@gsb.columbia.edu

In this web appendix, we describe our model, estimation procedure, and theoretical results in more detail. We also report more detailed results about our simulation studies and provide the aggregate data used in our empirical example. Web Appendix 1 provides a pseudocode summary of the data generating process assumed by our model specification from Section 2 of the main text. Web Appendix 2 describes the full procedure for computing the first two moments of the aggregate data, a key input to our proposed estimation procedure. Web Appendix 3 gives a derivation of the asymptotic properties of our proposed estimator, showing the conditions under which it is consistent and asymptotically normal and deriving the asymptotic distribution used to compute standard errors. Web Appendix 4 formalizes the identification conditions of our model and discusses identification of general heterogeneity distributions and selection functions. Web Appendix 5 presents detailed results of our simulation studies. Web Appendix 6 specifies the models we use as benchmarks and gives detailed performance comparisons. Lastly, Web Appendix 7 provides the ADD_q, $LOSS_q$, and END_q figures disclosed by Spotify that we use as the aggregate data in our empirical study.

Web Appendix 1: Data Generating Process

The full data generating process is described in Algorithm 1. Note that we can characterize the number of times individual i is re-acquired R_i as a geometric distribution because we assume that, each time an individual churns, there is a constant probability $\pi^{(RA)}$ that the individual will ever be re-acquired. As a result, the total number of re-acquisitions is geometrically distributed, independent of the duration of time between each re-acquisition.

Algorithm 1 Data generating process

for all $i = 1, 2, \ldots, N$ **do**

Draw i's parameter vector:

$$\log\left(\boldsymbol{\lambda}_i\right) \sim \mathcal{N}\left(\log\left(\boldsymbol{\lambda}_0\right), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\right)$$

Draw whether i is selected into the panel:

$$Z_i \sim Bernoulli\left(Logit^{-1}\left(\beta_0^{(Z)} + \left(\boldsymbol{\beta}^{(Z)}\right)' \log\left(\boldsymbol{\lambda}_i\right)\right)\right)$$

Draw whether i will ever be acquired:

$$A_i \sim Bernoulli\left(\pi^{(IA)}\right)$$

if $A_i = 1$ **then**

Draw initial acquisition time:

$$T_i^{(IA)} \sim Weibull\left(\lambda_i^{(IA)}, c^{(IA)}, \boldsymbol{\beta}^{(IA)}, \mathbf{x}_1, \mathbf{x}_2, \ldots\right)$$

Draw initial churn time:

$$T_i^{(IC)} \sim Weibull\left(\lambda_i^{(IC)}, c^{(IC)}, \boldsymbol{\beta}^{(IC)}, \mathbf{x}_1, \mathbf{x}_2, \ldots\right)$$

Draw the count of times i is to be re-acquired:

$$R_i \sim Geometric\left(\pi^{(RA)}\right)$$

if $R_i > 0$ **then**

for all $r = 1, 2, \ldots, R_i$ **do**

Draw re-acquisition time:

$$T_{ir}^{(RA)} \sim Weibull\left(\lambda_i^{(RA)}, c^{(RA)}, \boldsymbol{\beta}^{(RA)}, \mathbf{x}_1, \mathbf{x}_2, \ldots\right)$$

Draw churn time:

$$T_{ir}^{(RC)} \sim Weibull\left(\lambda_i^{(RC)}, c^{(RC)}, \boldsymbol{\beta}^{(RC)}, \mathbf{x}_1, \mathbf{x}_2, \ldots\right)$$

return $\left\{Z_i, T_i^{(IA)}, T_i^{(IC)}, T_{i1}^{(RA)}, T_{i1}^{(RC)}, T_{i2}^{(RA)}, T_{i2}^{(RC)}, \ldots, T_{iR_i}^{(RA)}, T_{iR_i}^{(RC)}\right\}$

Web Appendix 2: Derivation of Mean and Variance of Aggregate Data

In this web appendix, we derive the asymptotic distribution of the aggregated summary statistics $\mathbf{D}_N$ conditional on the panel data.[1] Recall that

$$\mathbf{Y}_i \coloneqq \left(IA_{i1}, \ldots, IA_{iM}, IC_{i1}, \ldots, IC_{iM}, RA_{i1}, \ldots, RA_{iM}, RC_{i1}, \ldots, RC_{iM}\right), \tag{1}$$

[1] As noted in Equation 8 in the main body of the paper, the observable data consists of the panel selection outcomes, the observed panel data, and the aggregate data ($z_{1:N}$, $\tilde{\mathbf{y}}_{1:N}$, and $\mathbf{d}$, respectively). As such, the joint likelihood of the data can be decomposed into the product of three terms, $P_{\boldsymbol{\theta}}\left(Z_{1:N} = z_{1:N}\right)$, $P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_{1:N} = \tilde{\mathbf{y}}_{1:N} | Z_{1:N} = z_{1:N}\right)$, and $P_{\boldsymbol{\theta}}\left(\mathbf{D}_N = \mathbf{d} | Z_{1:N} = z_{1:N}, \tilde{\mathbf{Y}}_{1:N} = \tilde{\mathbf{y}}_{1:N}\right)$. In this web appendix, we obtain the asymptotic distribution associated

where each element of $\mathbf{Y}_i$ is (marginally) a Bernoulli random variable. As discussed in Section 3.3, the proxy likelihood of the aggregated data $\mathbf{D}_N$ given the panel data can be computed by deriving the first two moments of $\mathbf{Y}_i$. Here, we provide a step-by-step procedure to compute these moments, then use them to obtain the asymptotic distribution of $\mathbf{D}_N$ given the panel data, through a series of six steps:

1. In Web Appendix 2.1, we derive all marginal probabilities that each of the elements of $\mathbf{Y}_i$ from Equation 1 are equal to 1 conditional upon $\boldsymbol{\lambda}$, $p_A(m|\boldsymbol{\lambda})$ for $A \in \{IA, IC, RA, RC\}$ and $m \in \{1, 2, \ldots, M\}$. We then derive all analogous pairwise joint probabilities $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$ for $(A, B) \in \{IA, IC, RA, RC\}^2$ and $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$. These derivations assume that all N population members become prospects at the beginning of commercial operations (i.e., at $m = 0$).

2. In Web Appendix 2.2, we use these probabilities to obtain the first two moments of $\mathbf{Y}_i$ conditional upon $\boldsymbol{\lambda}$, assuming prospects are all born in month 0.

3. In Web Appendix 2.3, we obtain the first two unconditional moments of $\mathbf{Y}_i$ by marginalizing out $\boldsymbol{\lambda}$, assuming prospects are all born in month 0.

4. In Web Appendix 2.4, we obtain the asymptotic distribution of total monthly initial and repeat acquisitions and losses given the panel data,

$$(\mathbf{IA}'\ \mathbf{IC}'\ \mathbf{RA}'\ \mathbf{RC}')' \coloneqq (IA_{\bullet 1}, \ldots, IA_{\bullet M}, IC_{\bullet 1}, \ldots, IC_{\bullet M}, RA_{\bullet 1}, \ldots, \ldots, RC_{\bullet M}),$$

where $IA_{\bullet m}$ represents total initial acquisitions in month m and $IC_{\bullet m}$, $RA_{\bullet m}$, and $RC_{\bullet m}$ are defined analogously, again assuming that prospects are all born in month 0.

5. In Web Appendix 2.5, we obtain the asymptotic distribution of $(\mathbf{IA}'\ \mathbf{IC}'\ \mathbf{RA}'\ \mathbf{RC}')'$ given the panel data after relaxing the assumption that all N population members become prospects in month 0, instead allowing for the possibility that N_m prospects are born in month m, $m \in \{0, 1, \ldots, M-1\}$.

6. In Web Appendix 2.6, we use the asymptotic distribution of $(\mathbf{IA}'\ \mathbf{IC}'\ \mathbf{RA}'\ \mathbf{RC}')'$ to obtain the asymptotic distribution of $\mathbf{D}_N$, which is parameterized by $\boldsymbol{\mu}_N(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_N(\boldsymbol{\theta})$, conditional on the panel data.

To provide further structure for the derivations that follow, the full procedures we use to calculate $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ are summarized in Algorithms 2 and 3, respectively.

In the sections that follow, we elaborate upon each of the steps summarized in the algorithms above.

2.1. $p_A(m|\boldsymbol{\lambda})$ and $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$

Note that in general, naively computing $p_A(m|\boldsymbol{\lambda})$ or $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$ would require summing up to $O(2^m)$ terms, since for the RA and RC processes, these probabilities require marginalizing over all possible sequences of acquisition and churn times prior to period m (analogously, all possible sequences of acquisition and churn times prior to m_1 and between m_1 and m_2 for $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$). However, through the use of belief propagation algorithms (Pearl 1988), we can compute these probabilities recursively, requiring only summing over $O(m)$ terms to compute $p_A(m|\boldsymbol{\lambda})$ and $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$. Our derivations below do not invoke any parametric assumptions, and so similar algorithms can be used for other models that have Markovian structure.

with the third term in this product, which is why we condition upon the panel data. Computing the unconditional asymptotic distribution of the aggregate data will "double count" the panel data in the joint likelihood expression.

Algorithm 2 Pseudocode for computing $\boldsymbol{\mu}(\boldsymbol{\theta})$

function CALCULATEMU($\boldsymbol{\theta}$, $z_{1:N}$, $\tilde{\mathbf{y}}_{1:N}$, $\mathbf{K}$, K)

for all $k = 1, 2, \ldots, K$ **do**

Using the k-th term of a 4-dimensional Halton sequence, simulate $\boldsymbol{\lambda}^{(k)}$ from the mixing distribution

$$\log\left(\boldsymbol{\lambda}^{(k)}\right) \sim \mathcal{N}\left(\log\left(\boldsymbol{\lambda}_0\right), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\right)$$

Compute all event probabilities for non-panel and inactive panel members given $\boldsymbol{\lambda}^{(k)}$, $p_A^{np}(m|\boldsymbol{\lambda}^{(k)})$ and $p_A^{pi}(m|\boldsymbol{\lambda}^{(k)})$, respectively, for $A \in \{IA, IC, RA, RC\}$ and $m \in \{1, 2, \ldots, M\}$, as derived in Web Appendix 2.1.

for all $m = 0, 1, \ldots, M-1$ **do**

Compute the number of non-panel members in the month m prospect pool, where $\mathbf{1}(A)$ is an indicator variable for event A, N_m is the size of prospect pool m, and N^p is the size of the panel population, recalling that all panel members come from the initial prospect pool:

$$N^{np}(m) \coloneqq N_m - \mathbf{1}(m = 0)N^p$$

for all $k = 1, 2, \ldots, K$ **do**

Using the event probabilities for non-panel members $p_A^{np}(m|\boldsymbol{\lambda}^{(k)})$, compute the first moment of $\mathbf{Y}$ for non-panel members from the month m prospect pool given $\boldsymbol{\lambda}^{(k)}$ (Equation 8):

$$EY_k^{np}(m) \coloneqq E(\mathbf{Y}|Z = 0, \boldsymbol{\lambda}^{(k)}, M = m)$$

if $m = 0$ **then**

Using the event probabilities for inactive panel members $p_A^{pi}(m|\boldsymbol{\lambda}^{(k)})$, compute the first moment of $\mathbf{Y}$ for inactive panel members given $\boldsymbol{\lambda}^{(k)}$ (Web Appendix 2.2):

$$EY_k^{pi} \coloneqq E(\mathbf{Y}|Z = 1, \tilde{Y} \neq \mathbf{Y}, \boldsymbol{\lambda}^{(k)}, M = m)$$

Marginalize $\boldsymbol{\lambda}$ out of the expressions for the first moment of $\mathbf{Y}$ for non-panel and, if $m = 0$, inactive panel members from the month m prospect pool (Web Appendix 2.3):

$$EY^{np}(m) \coloneqq \frac{1}{K}\sum_{k=1}^{K} EY_k^{np}(m)$$

$$EY^{pi} \coloneqq \frac{1}{P(Z = 1, \tilde{\mathbf{Y}}_i \neq \mathbf{Y}_i|\boldsymbol{\lambda}^{(k)})}\sum_{k=1}^{K} P(Z = 1, \tilde{\mathbf{Y}}_i \neq \mathbf{Y}_i|\boldsymbol{\lambda}^{(k)})EY_k^{pi} \quad \text{if } m = 0$$

Compute $\boldsymbol{\mu}_N$ by summing each prospect's contribution to the expected aggregate summary statistics – non-panel members (first term), active panel members (second) and inactive panel members (third), where N^{pi} is the number of inactive panel members (all from the month 0 prospect pool by assumption):

$$\boldsymbol{\mu}_N = \sum_{m=0}^{M} N^{np}(m)\mathbf{K}EY^{np}(m) + \sum_{i=1}^{N} \mathbf{K}\mathbf{y}_i\mathbf{1}((Z_i = 1)\&(Y_i = \tilde{Y}_i)) + N^{pi}\mathbf{K}EY^{pi}$$

return $\boldsymbol{\mu}_N$

First, we derive the marginal probabilities $p_A(m)$ for $A \in \{IA, IC, RA, RC\}$ and $m \in \{1, 2, \ldots, M\}$ recursively. Assume that all N population members are born as prospects in month 0 (an assumption we will relax in Web Appendix 2.5), so that prospects may be initially acquired starting in month 1. All probabilities not

Algorithm 3 Pseudocode for computing $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

function CALCULATESIGMA($\boldsymbol{\theta}$, $z_{1:N}$, $\tilde{\mathbf{y}}_{1:N}$, $\mathbf{K}$, K)

for all $k = 1, 2, \ldots, K$ **do**

Using the k-th term of a 4-dimensional Halton sequence, simulate $\boldsymbol{\lambda}^{(k)}$ from the mixing distribution

$$\log\left(\boldsymbol{\lambda}^{(k)}\right) \sim \mathcal{N}\left(\log\left(\boldsymbol{\lambda}_0\right), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\right)$$

Compute all joint event probabilities for non-panel and inactive panel members given $\boldsymbol{\lambda}^{(k)}$, $p^{np}_{A,B}(m_1, m_2|\boldsymbol{\lambda}^{(k)})$ and $p^{pi}_{A,B}(m_1, m_2|\boldsymbol{\lambda}^{(k)})$, respectively, for $(A, B) \in \{IA, IC, RA, RC\}^2$ and $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$, as derived in Web Appendix 2.1.

for all $m = 0, 1, \ldots, M-1$ **do**

Compute the number of non-panel members in the month m prospect pool, where $\mathbf{1}(A)$ is an indicator variable for event A, N_m is the size of prospect pool m, and N^p is the size of the panel population, recalling that all panel members come from the initial prospect pool:

$$N^{np}(m) \coloneqq N_m - \mathbf{1}(m = 0)N^p$$

for all $k = 1, 2, \ldots, K$ **do**

Using the joint event probabilities for non-panel members $p^{np}_{A,B}(m_1, m_2|\boldsymbol{\lambda}^{(k)})$, compute the second moment of $\mathbf{Y}$ for non-panel members given $\boldsymbol{\lambda}^{(k)}$ (Equation 9):

$$CovY^{np}_k(m) \coloneqq Cov(\mathbf{Y}|Z = 0, \boldsymbol{\lambda}^{(k)}, M = m)$$

if $m = 0$ **then**

Using the joint event probabilities for inactive panel members $p^{pi}_{A,B}(m_1, m_2|\boldsymbol{\lambda}^{(k)})$, compute the second moment of $\mathbf{Y}$ for inactive panel members given $\boldsymbol{\lambda}^{(k)}$ (Web Appendix 2.2):

$$CovY^{pi}_k \coloneqq Cov(\mathbf{Y}|Z = 1, \tilde{Y} \neq \mathbf{Y}, \boldsymbol{\lambda}^{(k)}, M = m)$$

Marginalize $\boldsymbol{\lambda}$ out of the expressions for the second moment of $\mathbf{Y}$ for non-panel and, if $m = 0$, inactive panel members from the month m prospect pool (Web Appendix 2.3):

$$CovY^{np}(m) \coloneqq \frac{1}{K}\sum_{k=1}^{K} CovY^{np}_k(m)$$

$$CovY^{pi} \coloneqq \frac{1}{P(Z = 1, \tilde{\mathbf{Y}}_i \neq \mathbf{Y}_i|\boldsymbol{\lambda}^{(k)})}\sum_{k=1}^{K} P(Z = 1, \tilde{\mathbf{Y}}_i \neq \mathbf{Y}_i|\boldsymbol{\lambda}^{(k)})CovY^{pi}_k \quad \text{if } m = 0$$

Compute $\boldsymbol{\Sigma}_N$ by summing each prospect's contribution to the covariance of the aggregate summary statistics – non-panel members (first term) and inactive panel members (second), where N^{pi} is the number of inactive panel members (all from the month 0 prospect pool by assumption):

$$\boldsymbol{\Sigma}_N = \sum_{m=0}^{M} N^{np}(m)\mathbf{K}CovY^{np}(m)\mathbf{K}' + N^{pi}\mathbf{K}CovY^{pi}\mathbf{K}'$$

return $\boldsymbol{\Sigma}_N$

explicitly defined below are equal to 0 (e.g., if a probability is defined for $m \geq 2$, then the corresponding probability for $m = 1$ equals 0).

Let $p_{IA}(m')$ be the probability that a population member is initially acquired in month m'. Then conditional upon the rate parameters $\boldsymbol{\lambda} = [\lambda^{(IA)}, \lambda^{(IC)}, \lambda^{(RA)}, \lambda^{(RC)}]$ and suppressing fixed effect coefficients throughout ($c^{(A)}$ and $\boldsymbol{\beta}^{(A)}$ $\forall$ $A \in \{IA, IC, RA, RC\}$), the marginal probabilities associated with the initial acquisition process conditional upon $\boldsymbol{\lambda}$ are

$$p_{IA}(m'|\boldsymbol{\lambda}) = S^{(IA)}\left(m'-1|\lambda^{(IA)}\right) - S^{(IA)}\left(m'|\lambda^{(IA)}\right), \qquad m' \geq 1$$

where $S^{(IA)}(m)$ is defined as the survival probability of not having been initially acquired m months after having become a prospect, which is directly computable from the proportional hazards Weibull model.

Let $p_{IC}(m'')$ be the probability that a population member initially churns in month m''. As with the initial acquisition process, the marginal initial churn probabilities conditional upon $\boldsymbol{\lambda}$ are

$$p_{IC}(m''|\boldsymbol{\lambda}) = \sum_{m'=0}^{m''-1} p_{IC}(m''|m', \lambda^{(IC)}) p_{IA}(m'|\lambda^{(IA)}), \quad m'' \geq 2 \tag{2}$$

where $p_{IC}(m''|m', \lambda^{(IC)})$ is the probability that a customer churns in month m'', conditional upon their having been initially acquired in month m':

$$p_{IC}(m''|m', \lambda^{(IC)}) \coloneqq S^{(IC)}\left(m''-m'-1|\lambda^{(IC)}\right) - S^{(IC)}\left(m''-m'|\lambda^{(IC)}\right), \quad m'' > m', \; m'' \geq 2 \tag{3}$$

Let $p_{RA}(m')$ be the probability that a population member is reacquired in month m' after having churned in a previous month. Taking into account the possibility that the previous churn event could have been an initial churn or a repeat churn, the marginal repeat acquisition probabilities conditional upon $\boldsymbol{\lambda}$ are equal to

$$p_{RA}(m'|\boldsymbol{\lambda}) = \sum_{m''=2}^{m'-2} p_{RA}(m'|m'', \lambda^{(RA)}) \left(p_{IC}(m''|\lambda^{(IC)}) + p_{RC}(m''|\lambda^{(RC)})\right), \quad m' \geq 4 \tag{4}$$

$p_{RA}(m'|m'', \lambda^{(RC)})$ is the probability that a customer is re-acquired in month m', conditional upon their having churned (initial or repeat) in month m'', and is defined through difference in CDF's in the same way that $p_{IC}(m''|m', \lambda^{(IC)})$ is defined in Equation 3. Note that $p_{RA}(m'|\boldsymbol{\lambda}) = 0$ for $m' \leq 3$, because the earliest month in which a customer could churn is $m = 2$ (immediately after being acquired in $m = 1$), in which case they would be "re-born" as a prospect in month 3 and thus can first be reacquired in month 4, as described in point 5 of Appendix A in the main text.

While the first term in Equation 4 comes directly from a Weibull model probability and the second term was derived in Equation 2, the last term, $p_{RC}(m''|\lambda^{(RC)})$, is defined through a recursion for $m'' \geq 5$ (it is equal to 0 for $m'' \in \{1, 2, 3, 4\}$). Because $p_{RC}(m''|\lambda^{(RC)})$ (and thus $p_{RA}(m'|\lambda^{(RA)})$) are defined recursively, all terms corresponding to a given month m must be computed before computing the terms corresponding to month $m+1$. All previous periods' recursive probabilities are known when computing subsequent periods' probabilities (e.g., $p_{RA}(4|\lambda^{(RA)})$ is a function of $p_{RC}(2|\lambda^{(RC)})$, $p_{RC}(5|\lambda^{(RC)})$ is a function of $p_{RA}(4|\lambda^{(RA)})$ as we will see next, and so on).

Let $p_{RC}(m'')$ be the probability that a population member repeat churns in month m'' after having been repeat acquired in a previous month. This is essentially identical to the initial churn computation in Equation 2 except that it references the model probabilities for the repeat churn process and the recursively computed marginal probabilities for the repeat acquisition process instead of the initial acquisition process:

$$p_{RC}(m''|\boldsymbol{\lambda}) = \sum_{m'=3}^{m''-1} p_{RC}(m''|m', \lambda^{(RC)}) p_{RA}(m'|\lambda^{(RA)}), \quad m' \geq 5 \tag{5}$$

This completes the derivations for the marginal probabilities for the initial and repeat acquisition and churn processes conditional upon $\boldsymbol{\lambda}$. As noted above, all marginal probabilities can be computed in a single loop over all time periods $m = 1, 2, \ldots$. Next, we derive the 11 joint probabilities for any pair of events.

$p_{IA,IA}(m_1, m_2|\boldsymbol{\lambda})$ and $p_{IC,IC}(m_1, m_2|\boldsymbol{\lambda})$ are trivially equal to 0 for all $m_1 \neq m_2$, since by definition any individual can only be initially acquired or initially lost one time. This leaves the following nine joint probabilities to be computed. In all expressions that follow, without loss of generality, we assume that $m_2 > m_1$: when $m_2 < m_1$, we can just flip the arguments; when $m_2 = m_1$ the probability is trivially zero for $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$ when $A \neq B$ since a customer cannot, for instance, adopt and churn in the same month, and when $A = B$ the probability simply reduces to the marginal probability $p_A(m_1)$. As was the case with the marginal probabilities above, all probabilities not explicitly defined are equal to 0.

The joint probability of initial acquisition and initial loss follows directly from the model probabilities without any recursively defined terms:

$$p_{IA,IC}(m_1, m_2|\boldsymbol{\lambda}) = p_{IC}(m_2|m_1, \lambda^{(IC)}) p_{IA}(m_1|\lambda^{(IA)}), \; m_1 \geq 1, \; m_2 \geq m_1$$

The joint probability of initial acquisition and repeat acquisition is equal to the sum of two terms:

$$\begin{aligned} p_{IA,RA}(m_1, m_2|\boldsymbol{\lambda}) &= C_1 + C_2, \\ C_1 &= \sum_{m^*=m_1+1}^{m_2-2} p_{RA}(m_2|m^*, \lambda^{(RA)}) p_{IL}(m^*|m_1, \lambda^{(IC)}) p_{IA}(m_1, \lambda^{(IA)}), \\ &\qquad m_1 = 1, \ldots, M-3, \; m_2 \geq m_1 + 3 \\ C_2 &= \sum_{m^*=m_1+4}^{m_2-2} p_{RA}(m_2|m^*, \lambda^{(RA)}) p_{IA,RC}(m_1, m^*|\lambda^{(IA)}, \lambda^{(RC)}), \\ &\qquad m_1 = 2, \ldots, M-6, \; m_2 \geq m_1 + 6 \end{aligned} \tag{6}$$

C_1 and C_2 account for two possible scenarios. The first scenario is that m_2 is the customer's second acquisition time, while the second scenario is that m_2 is the time of the customer's third or higher acquisition time. In the former case, we only need to perform a one-fold convolution over the customer's initial churn time. In the latter case, rather than convoluting over all possible acquisition and churn paths, we only convolute over the most recent churn time, exploiting the memorylessness of the reacquisition process conditional upon the last time the customer had churned. $p_{IA,RC}(m_1, m^*|\lambda^{(IA)}, \lambda^{(RC)})$ is itself defined through a recursion, which we will specify in the derivations that follow.

The joint probability of initial acquisition and repeat churn is computed recursively, summing over all the possible months m^* that the customer was last reacquired:

$$p_{IA,RC}(m_1, m_2|\boldsymbol{\lambda}) = \sum_{m^*=m_1+3}^{m_2-1} p_{RC}(m_2|m^*, \lambda^{(RC)}) p_{IA,RA}(m_1, m^*|\lambda^{(IA)}, \lambda^{(RA)}),$$
$$m_1 = 1, 2, \ldots, M-4,\ m_2 \geq m_1 + 4 \tag{7}$$

The joint probability of initial churn and repeat acquisition is computed recursively, and is equal to the sum of two terms:

$$\begin{aligned} p_{IC,RA}(m_1, m_2|\boldsymbol{\lambda}) &= C_3 + C_4, \\ C_3 &= p_{RA}(m_2|m_1, \lambda^{(RA)}) p_{IC}(m_1|\lambda^{(IC)}),\ m_1 = 2, \ldots, M-2,\ m_2 \geq m_1 + 2 \\ C_4 &= \sum_{m^*=m_1+2}^{m_2-1} p_{RC}(m_2|m^*, \lambda^{(RC)}) p_{IC,RC}(m_1, m^*|\lambda^{(IC)}, \lambda^{(RC)}), \\ &\qquad m_1 = 2, 3, \ldots, M-5,\ m_2 \geq m_1 + 5 \end{aligned}$$

As in Equation 6, C_3 and C_4 account for two possible scenarios, the first being that m_2 was preceded by an initial churn, the second being that m_2 was preceded by a repeat churn. As before, we use a recursive term in C_4 to avoid having to convolute over all possible paths between the initial churn in month m_1 and the repeat churn in month m^*.

The joint probability of initial churn and repeat churn sums over all the possible months m^* which repeat acquisition could have happened in:

$$p_{IC,RC}(m_1, m_2) = \sum_{m^*=m_1+2}^{m_2-1} p_{RC}(m_2|m^*, \lambda^{(RC)}) p_{IC,RA}(m_1, m^*|\lambda^{(IC)}, \lambda^{(RA)})$$
$$m_1 = 2, 3, \ldots, M-3,\ m_2 \geq m_1 + 3$$

The joint probability of repeat acquisition in month m_1 and repeat acquisition in month m_2 sums over all possible months m^* in which the last repeat churn event prior to m_2 had occurred:

$$p_{RA,RA}(m_1, m_2|\boldsymbol{\lambda}) = \sum_{m^*=m_1+1}^{m_2-2} p_{RA}(m_2|m^*, \lambda^{(RA)}) p_{RA,RC}(m_1, m^*|\lambda^{(RA)}, \lambda^{(RC)}),$$
$$m_1 = 4, \ldots, M-3,\ m_2 \geq m_1 + 3$$

The joint probability of repeat acquisition in month m_1 and repeat churn in month m_2 is the sum of two terms, accounting for the possibilities that (1) there were no repeat churn events and (2) there was one or more repeat churn event event between months m_1 and m_2:

$$\begin{aligned} p_{RA,RC}(m_1, m_2|\boldsymbol{\lambda}) &= C_5 + C_6, \\ C_5 &= p_{RA}(m_1|\lambda^{(RA)}) p_{RC}(m_2|m_1, \lambda^{(RC)}),\ m_1 = 4, \ldots, M-1,\ m_2 \geq m_1 + 1 \\ C_6 &= \sum_{m^*=m_1+3}^{m_2-1} p_{RC}(m_2|m^*, \lambda^{(RC)}) p_{RA,RA}(m_1, m^*|\lambda^{(RA)}), \\ &\qquad m_1 = 4, \ldots, M-4,\ m_2 \geq m_1 + 4 \end{aligned}$$

Algorithm 4 Pseudocode for Marginal and Joint Probability Recursion

for all $m_1 = 1, 2, \ldots, M$ **do**
  Compute $p_A(m_1|\boldsymbol{\lambda})$ for $A \in \{IA, IC, RA, RC\}$
  for all $m_2 = m_1 + 1, m_1 + 2, \ldots, M$ **do**
    Compute $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$ for $(A, B) \in \{IA, IC, RA, RC\}^2$
return $p_A(m_1|\boldsymbol{\lambda})$ and $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$ for $(A, B) \in \{IA, IC, RA, RC\}^2$
and $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$

Note that in both cases, m_1, the first month in which a customer may be reacquired, can be no less than 4, as a customer can be acquired no earlier than month 1, initially churn no earlier than month 2, be "reborn" as a prospect in month 3, then be reacquired no earlier than month 4.

The joint probability of repeat churn in month m_1 and repeat acquisition in month m_2 is the sum of two terms, accounting for whether (1) there were no intervening repeat churn events or (2) there were one or more intervening repeat churn events between months m_1 and m_2:

$$\begin{aligned} p_{RC,RA}(m_1, m_2|\boldsymbol{\lambda}) &= C_7 + C_8, \\ C_7 &= p_{RC}(m_1|\lambda^{(RC)}) p_{RA}(m_2|m_1, \lambda^{(RA)}),\ m_1 = 5, \ldots, M-2,\ m_2 \geq m_1 + 2 \\ C_8 &= \sum_{m^*=m_1+3}^{m_2-2} p_{RA}(m_2|m^*, \lambda^{(RA)}) p_{RC,RC}(m_1, m^*|\lambda^{(RC)}), \\ &\qquad m_1 = 5, \ldots, M-5,\ m_2 \geq m_1 + 5 \end{aligned}$$

The joint probability of repeat churn in month m_1 and repeat churn in month m_2 sums over all possible months in which the final repeat acquisition month m^* prior to m_2 had occurred:

$$\begin{aligned} p_{RC,RC}(m_1, m_2) = \sum_{m^*=m_1+2}^{m_2-1} & p_{RC}(m_2|m^*, \lambda^{(RC)}) p_{RC,RA}(m_1, m^*|\lambda^{(RC)}, \lambda^{(RA)}) \\ & m_1 = 5, \ldots, M-3,\ m_2 \geq m_1 + 3 \end{aligned}$$

This completes all the formulas needed to compute $p_A(m_1|\boldsymbol{\lambda})$ and $p_{A,B}(m_1, m_2|\boldsymbol{\lambda})$ for $(A, B) \in \{IA, IC, RA, RC\}^2$ and $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$. All marginal probability calculations in month m_1 are a function of marginal probabilities from previous months $1, 2, \ldots, m_1 - 1$. The same is true for all joint probability calculations. Therefore, we successively evaluate these terms at increasing values of (m_1, m_2). Pseudocode for the recursion is shown in Algorithm 4. Note that, since the marginal probabilities are only a function of other marginal probabilities, if we only need to compute marginal probabilities then we can skip the inner loop of Algorithm 4 to reduce the computational burden. This is useful when we compute $\boldsymbol{\mu}_N(\boldsymbol{\theta})$, which we will show below is only a function of marginal probabilities.

A final subtlety is that the moment equations defined in Equation 9 of the main text are conditioned on the panel selection outcome Z_i and observable panel data $\tilde{\mathbf{Y}}_i$, so as not to "double count" the panel data. For this reason, we need to compute the above probabilities conditional on Z_i and $\tilde{\mathbf{Y}}_i$.

For the non-panel population, this is simple: by assumption, Z_i is conditionally independent of $\mathbf{Y}_i$ given heterogeneity parameter $\boldsymbol{\lambda}_i$ (Section 2.3 of the main text), and by definition there is no observable panel data for the non-panel population, i.e. $\tilde{\mathbf{Y}}_i = \emptyset$ (Section 3.1 of the main text). Therefore, when $Z_i = 0$, Z_i and $\tilde{\mathbf{Y}}_i$ are uninformative about $\mathbf{Y}_i|\boldsymbol{\lambda}_i$ and we can simply drop the conditioning and use the formulas above. We denote the collection of marginal and joint probabilities for non-panel members by $p_A^{np}(m_1|\boldsymbol{\lambda})$ and $p_{A,B}^{np}(m_1, m_2|\boldsymbol{\lambda})$, respectively.

For panel members, i.e. those for whom $Z_i = 1$, the probability computations must be modified to condition on the observed panel data $\tilde{\mathbf{Y}}_i$. In general, when some but not all elements of $\mathbf{Y}_i$ are observed in $\tilde{\mathbf{Y}}_i$ for panel members, the above recursive equations still apply, but with the probabilities of observed elements replaced by their observed values and the other probabilities rescaled accordingly. In our empirical application, as discussed in points 7 and 8 of Appendix A of the main text, there are two types of panel members: those who were initially acquired by Spotify during the panel observation period between January 2015 and September 2018 ("panel actives"), for whom we observe their full vector of behavior, i.e. $\tilde{\mathbf{Y}}_i = \mathbf{Y}_i$; and those who were not ("panel inactives"), for whom we only observe that $IA_{im} = 0$ over the months m corresponding to the panel observation period. Thus, for panel actives, we observe their full outcome vector and so their outcomes are deterministic conditional on the panel data and no recursion is required; all probabilities are replaced by the observed binary outcomes. Conversely, for panel inactives, we just know that $IA_{im} = 0$ during the panel observation period; thus, we apply the same recursion above, but condition on this information by zeroing out $p_{IA}(m|\boldsymbol{\lambda})$ over the months corresponding to the panel observation period and then rescaling all the remaining non-zero $p_{IA}(m|\boldsymbol{\lambda})$ terms to sum to one. We denote the collection of marginal and joint probabilities for panel inactives by $p_A^{pi}(m_1|\boldsymbol{\lambda})$ and $p_{A,B}^{pi}(m_1, m_2|\boldsymbol{\lambda})$, respectively.

2.2. First and Second Conditional Moments of $\mathbf{Y}_i$

We separately consider the first two moments of the elements of $\mathbf{Y}_i$ conditional upon $\boldsymbol{\lambda}$ given the observed panel data $\tilde{\mathbf{Y}}_i$ for non-panel members, active panel members, and inactive panel members. The moments for non-panel and inactive panel members follow trivially from the marginal and joint probability expressions we derived in Section 2.1. Since the elements of $\mathbf{Y}_i$ are all Bernoulli random variables, the moments for non-panel members are simply:

$$E(A_{im_1}|\boldsymbol{\lambda}, \tilde{\mathbf{Y}}_i = \emptyset) = E(A_{m_1}|\boldsymbol{\lambda}, \tilde{\mathbf{Y}}_i = \emptyset) = p_A^{np}(m_1|\boldsymbol{\lambda}) \tag{8}$$

$$\begin{aligned} Cov(A_{im_1}, B_{im_2}|\boldsymbol{\lambda}, \tilde{\mathbf{Y}}_i = \emptyset) &= Cov(A_{m_1}, B_{m_2}|\boldsymbol{\lambda}, \tilde{\mathbf{Y}}_i = \emptyset) \\ &= p_{A,B}^{np}(m_1, m_2|\boldsymbol{\lambda}) - p_A^{np}(m_1|\boldsymbol{\lambda}) p_B^{np}(m_2|\boldsymbol{\lambda}) \end{aligned} \tag{9}$$

for $(A, B) \in \{IA, IC, RA, RC\}$. The analogous moment expressions for the panel inactives follow naturally, replacing $p_A^{np}(m_1|\boldsymbol{\lambda})$ with $p_A^{pi}(m_1|\boldsymbol{\lambda})$ and $p_{A,B}^{np}(m_1, m_2|\boldsymbol{\lambda})$ with $p_{A,B}^{pi}(m_1, m_2|\boldsymbol{\lambda})$ in the expressions above.

For active panel members, their behaviors are deterministic given the panel data, so their expectations are simply their observed behaviors, and the corresponding variances and covariances are zero. In practice,

when computing the aggregate moments, we simply omit the panel actives and subtract their observed panel activity off of $\mathbf{D}_N$ to compensate; this results in an exactly equivalent proxy likelihood to if we included the panel actives explicitly, since it just shifts the distribution of $\mathbf{D}_N$ by a deterministic constant.

2.3. Moving from Conditional to Marginal Moments of $\mathbf{Y}_i$

The expressions above assume knowledge of a particular individual's rate parameters $\boldsymbol{\lambda}$, which we must integrate out to obtain the corresponding unconditional moment expressions. We approximate the integrals using draws from a Halton sequence. If we were not conditioning on the panel selection outcome Z_i and panel data $\tilde{\mathbf{Y}}_i$, we could easily perform this marginalization by taking a simple average across sampled $\boldsymbol{\lambda}$ vectors obtained using Halton draws from the heterogeneity distribution $g(\boldsymbol{\lambda})$ (i.e., from $\mathcal{N}(\log(\boldsymbol{\lambda}_0), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}})$). That is, indexing the Halton draws by $k \in \{1, 2, \ldots K\}$,

$$\begin{aligned} E(A_{m_1}) &= \int_{\mathbb{R}^4_+} E(A_{m_1}|\boldsymbol{\lambda}) g(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \approx \frac{1}{K} \sum_{k=1}^{K} E(A_{m_1}|\boldsymbol{\lambda}^{(k)}) \\ Cov(A_{m_1}, B_{m_2}) &= \int_{\mathbb{R}^4_+} Cov(A_{m_1}, B_{m_2}|\boldsymbol{\lambda}) g(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \approx \frac{1}{K} \sum_{k=1}^{K} Cov(A_{m_1}, B_{m_2}|\boldsymbol{\lambda}^{(k)}) \end{aligned} \tag{10}$$

where $\boldsymbol{\lambda}^{(k)}$ is a Halton draw from $g(\boldsymbol{\lambda})$, so for all population members, the mean vector associated with $\mathbf{Y}$ would simple be:

$$E(\mathbf{Y}) \approx \frac{1}{K} \sum_{k=1}^{K} E[\mathbf{Y}|\boldsymbol{\lambda}^{(k)}]$$

where each conditional mean vector in the summand has its elements populated according to the event-specific conditional means shown in Equation 8, with a corresponding covariance matrix $Cov(\mathbf{Y})$ obtained using the covariance terms from Equation 9.

Conditioning on the panel selection outcome Z_i and observable panel data $\tilde{\mathbf{Y}}_i$, however, complicates the marginalization procedure, since we must integrate over the posterior mixing distribution $g\left(\boldsymbol{\lambda}_i|Z_i, \tilde{\mathbf{Y}}_i\right)$ (in addition to modifying the probability computations as described above in Web Appendix 2.1). This posterior is not analytically tractable to compute, but can be approximated easily using importance sampling. By Bayes' theorem,

$$g\left(\boldsymbol{\lambda}_i|Z_i, \tilde{\mathbf{Y}}_i\right) \propto g(\boldsymbol{\lambda}_i) \cdot P\left(Z_i|\boldsymbol{\lambda}_i\right) \cdot P\left(\tilde{\mathbf{Y}}_i|\boldsymbol{\lambda}_i, Z_i\right). \tag{11}$$

For the non-panel members, $\tilde{\mathbf{Y}}_i$ is a degenerate random variable, and hence this formula simplifies to:

$$g\left(\boldsymbol{\lambda}_i|Z_i = 0, \tilde{\mathbf{Y}}_i\right) = g\left(\boldsymbol{\lambda}_i|Z_i = 0\right) \propto g(\boldsymbol{\lambda}_i) \cdot P\left(Z_i = 0|\boldsymbol{\lambda}_i\right)$$

and so, instead of taking a simple average of the Halton draws $\boldsymbol{\lambda}^{(k)}$ from $g(\boldsymbol{\lambda})$ as we would if not conditioning on the panel data, we instead take a weighted average using importance sampling, where the importance weights are the probabilities of panel non-selection. That is, indexing the Halton draws by $k \in \{1, 2, \ldots K\}$ as before,

$$\begin{aligned} E(A_{m_1}|Z_i = 0, \tilde{\mathbf{Y}}_i) &\approx \frac{1}{\sum_{k=1}^{K} w_k^{np}} \sum_{k=1}^{K} w_k^{np} E(A_{m_1}|\boldsymbol{\lambda}^{(k)}) \\ Cov(A_{m_1}, B_{m_2}|Z_i = 0, \tilde{\mathbf{Y}}_i) &\approx \frac{1}{\sum_{k=1}^{K} w_k^{np}} \sum_{k=1}^{K} w_k^{np} Cov(A_{m_1}, B_{m_2}|\boldsymbol{\lambda}^{(k)}) \end{aligned} \tag{12}$$

where $\boldsymbol{\lambda}^{(k)}$ is defined as before and w_k^{np} is the probability that a customer with rate parameters $\boldsymbol{\lambda}^{(k)}$ is not selected into the panel (using the logit selection formula from Equation 4 of the main text), relative to the unconditional probability that a population member is selected into the panel:

$$w_k^{np} = 1 - P(Z_i = 1|\boldsymbol{\lambda}^{(k)}).$$

For panel inactives, $\tilde{\mathbf{Y}}_i$ is not a degenerate random variable and so the third term in Equation 11 does not drop off. Accordingly, for panel inactives, we can derive their unconditional moments using Equation 12, but replacing the importance weights from w_k^{np} by:

$$w_k^{pi} = P(Z_i = 1|\boldsymbol{\lambda}^{(k)}) \cdot P(\tilde{\mathbf{Y}}_i \neq \mathbf{Y}_i|Z_i = 1, \boldsymbol{\lambda}^{(k)})$$

with $P(\tilde{\mathbf{Y}}_i \neq \mathbf{Y}_i|Z_i = 1, \boldsymbol{\lambda}^{(k)})$ computed according to the conditions described at the end of Web Appendix 2.1.

For panel actives, marginalization is unnecessary since as discussed above their activity is deterministic conditional on the panel data. Thus, using the above importance sampling formulas give us all we need to compute the marginal moments of $\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i$; for non-panel members and panel inactives, we use importance sampling for marginalization, while for panel actives we simply replace their expectation terms by their observed activity and their variance and covariance terms by zero.

As an aside, as mentioned in Section 3.3, computing the panel data likelihood (the second term of the proxy likelihood function given in Equation 10 of the main text) requires marginalizing over the posterior $g(\boldsymbol{\lambda}_i|Z_i = 1)$. A similar importance weighting scheme can be used to compute this term. The marginal likelihood of activity by panel actives is the corresponding weighted average of sampled conditional likelihoods:

$$P\left(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i|Z_i = z_i\right) \approx \frac{1}{\sum_{k=1}^{K} w_k^{pa}} \sum_{k=1}^{K} w_k^{pa} P(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i|\boldsymbol{\lambda}^{(k)}),$$

where $\boldsymbol{\lambda}^{(k)}$ is defined as before and

$$w_k^{pa} = P(Z_i = 1|\boldsymbol{\lambda}^{(k)}),$$

which is precisely the formula used to compute the marginal panel likelihood in Algorithm 1 of the main text. This expression is only used when calculating the likelihood of the panel data, which can be computed exactly (conditional on $\boldsymbol{\lambda}_i$) as it is a series of simple Weibull timing events. As noted earlier, when computing the moments of the aggregate data, the activity of panel actives is deterministic, obviating the need for marginalization when computing their contribution to the aggregate moments.

2.4. Asymptotic Distribution of (IA′ IC′ RA′ RC′)′

The asymptotic distribution of total initial and repeat acquisitions and losses summing over all population members $\mathbf{Y}_i$, [**IA IC RA RC**] given their selection outcomes, follows immediately from the results of the previous section. The random vectors $\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i$, for $i = 1, 2, \ldots, N$, are independently distributed (though non-identically distributed, due to conditioning on the panel data) with mean vectors $E(\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i)$ and covariance matrices $Cov(\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i)$, so by Eicker's multivariate central limit theorem for non-identically distributed random vectors (Eicker 1967),

$$\frac{1}{\sqrt{N}} \left(\sum_{i=1}^{N} Y_i - E(\mathbf{Y}_i|Z_i, \tilde{\mathbf{Y}}_i) \right) \rightarrow_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{Y}}) \tag{13}$$

where $\mathbf{0}$ represents a $4M$-length vector of 0's and

$$\mathbf{\Sigma_Y} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} Cov(\mathbf{Y}_i | Z_i, \tilde{\mathbf{Y}}_i).$$

2.5. Multiple Prospect Pools

The preceding results assume that all N population members become prospects in month $m = 0$, and that no population members will become prospects in future months $m = 1, 2, \ldots$. This is unlikely to be true in real-world settings, due to factors such as household formation, population growth, and corporate decisions to enter new geographies over time. It is more realistic to assume instead that these N population members become prospects throughout the observation period. Let

$$\mathbf{N} \coloneqq [N_0 \; N_1 \; \cdots \; N_{M-1}]$$

represent the number of prospects born in each month $m = 0, 1, \ldots, M-1$. If all prospects were born in month 0, then $\mathbf{N} = (N, 0, \ldots, 0)$.

While population members born as prospects in different months, collectively referred to as "prospect pools," are assumed to share the same underlying Weibull timing models for initial and repeat acquisition and churn (e.g., the duration of time from when population members become prospects to when they are initially acquired are stochastically the same across prospect pools), they will be exposed to the time-varying covariates $\mathbf{X}_{1:M}$ according to different schedules (e.g. members "born" in month 0 will have their probabilities augmented by $\mathbf{X}_3$ in the 3rd month after becoming prospects, while those "born" in month 2 will have their probabilities augmented by $\mathbf{X}_3$ in the 1st month after becoming prospects). As such, the mean vectors and covariance matrices $E(\mathbf{Y}_i)$ and $Cov(\mathbf{Y}_i)$ across "prospect pools" are not simply lagged versions of each other. For this reason, $E(\mathbf{Y}_i)$ and $Cov(\mathbf{Y}_i)$ must be computed separately for each prospect pool using the recursive calculation outlined in Algorithm 4. Additionally, as mentioned in point 1 of Appendix A of the main text, all panel members are assumed to come from the initial prospect pool. Thus, the probability of panel selection is zero for all prospects born in months $m \geq 1$, and so for these prospect pools, marginalization can just be done through simple averaging as in Equation 10 since we do not need to condition on the panel outcome or panel data, and the $\mathbf{Y}_i$s are i.i.d. within prospect pools.

With the addition of prospect pools, the asymptotic distribution of $(\mathbf{IA}'\ \mathbf{IC}'\ \mathbf{RA}'\ \mathbf{RC}')'$ in Equation 13 remains the same, but $E(\mathbf{Y}|Z_i, \tilde{\mathbf{Y}}_i)$ and $Cov(\mathbf{Y}|Z_i, \tilde{\mathbf{Y}}_i)$ are replaced by $E(\mathbf{Y}|Z_i, \tilde{\mathbf{Y}}_i, M_i)$ and $Cov(\mathbf{Y}|Z_i, \tilde{\mathbf{Y}}_i, M_i)$ to also condition on the prospect pool M_i in which prospect i was born.

2.6. Asymptotic Distribution of $\mathbf{D}_N$

The summary statistics END_q, ADD_q, and $LOSS_q$ that we consider are all simple linear transformations of $(\mathbf{IA}'\ \mathbf{IC}'\ \mathbf{RA}'\ \mathbf{RC}')'$, and so the function s can be encoded by a suitable linear transformation matrix $\mathbf{K}$,

whose number of rows is equal to the number of summary statistics disclosed and whose number of columns is equal to $4M$. This follows immediately from the fact that

$$\begin{aligned} ADD_q &= \sum_{m=3(q-1)+1}^{3q} IA_{\bullet m} + RA_{\bullet m} \\ LOSS_q &= \sum_{m=3(q-1)+1}^{3q} IC_{\bullet m} + RC_{\bullet m} \\ END_q &= \sum_{m=1}^{3q} IA_{\bullet m} + RA_{\bullet m} - IC_{\bullet m} - RC_{\bullet m} \end{aligned} \tag{14}$$

It follows that $\mathbf{K}$ is in general defined such that

$$\mathbf{D}_N = \mathbf{K} \begin{pmatrix} \mathbf{IA} \\ \mathbf{IC} \\ \mathbf{RA} \\ \mathbf{RC} \end{pmatrix}. \tag{15}$$

As a result, the asymptotic distribution of the summary statistics $\mathbf{D}_N$ is equal to

$$\frac{1}{\sqrt{N}} \left(\mathbf{D}_N - \mathbf{K} \left(\sum_{i=1}^{N} E(\mathbf{Y}_i | Z_i, \tilde{\mathbf{Y}}_i, M_i) \right) \right) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{K} \boldsymbol{\Sigma}_{\mathbf{Y}} \mathbf{K}') \tag{16}$$

Accordingly, the asymptotic approximation to the distribution of $\mathbf{D}_N$ we use for computing the proxy likelihood is a multivariate normal distribution with mean vector

$$\boldsymbol{\mu}_N(\boldsymbol{\theta}) = \mathbf{K} \left(\sum_{i=1}^{N} E_{\boldsymbol{\theta}}(\mathbf{Y}_i | Z_i, \tilde{\mathbf{Y}}_i, M_i) \right)$$

and covariance matrix

$$\boldsymbol{\Sigma}_N(\boldsymbol{\theta}) = \mathbf{K} \left(\sum_{i=1}^{N} Cov_{\boldsymbol{\theta}}(\mathbf{Y}_i | Z_i, \tilde{\mathbf{Y}}_i, M_i) \right) \mathbf{K}',$$

where the subscript $\boldsymbol{\theta}$ indicates that the theoretical moments are computed at the parameter value $\boldsymbol{\theta}$. In general, the asymptotic distribution for any vector of aggregate moments which are affine transformations of $(\mathbf{IA}'\ \mathbf{IC}'\ \mathbf{RA}'\ \mathbf{RC}')'$ can be derived analogously.

Web Appendix 3: Asymptotic Properties of Proposed Estimator

In this section, we establish the asymptotic properties of our proposed estimator under some regularity conditions. First, in Section 3.1 we define general notation for the aggregate-disaggregate data fusion problem and the maximum proxy likelihood estimator; then, in Section 3.2 we establish the consistency of our estimator, and in Section 3.3 we establish asymptotic normality and derive the variance of the limiting distribution.

3.1. General Notation

Suppose we wish to estimate a model for a target population of size N, with each member of the population represented by a random vector $(\mathbf{Y}_i', Z_i)' \in \mathcal{Y} \times \{0,1\}$ with $\mathcal{Y} \subseteq \mathbb{R}^k$ for $i = 1, 2, \ldots, N$, and $(\mathbf{Y}_i', Z_i)'$ independent and identically distributed. $\mathbf{Y}_i$ represents the behavior(s) of interest for population member i, while the binary variable Z_i represents whether i is in the population cross-section for which granular data is observed. Z_i is observed for all $i = 1, 2, \ldots, N$, with $Z_i = 1$ denoting that granular data is observed for population member i, while $Z_i = 0$ denotes that they were not. We assume that N is known.

In our application, N represents the total number of "prospects" (i.e., individuals who may be acquired as customers at some point in the future) a company has. The behaviors of interest that we model for these prospects, encoded by $\mathbf{Y}_i$, are their acquisition and churn decisions each period, while Z_i indicates whether they are in the credit card panel or not.

In general, consider a model consisting of a class of probability distributions indexed by parameter vector $\boldsymbol{\theta}$, denoting the parameter space by Θ:

$$\begin{pmatrix} \mathbf{Y}_i \\ Z_i \end{pmatrix} \sim F_{\boldsymbol{\theta}} \ \{\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\} \tag{17}$$

Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0 \in \Theta$.

Define the random variable $\tilde{\mathbf{Y}}_i$ as:

$$\tilde{\mathbf{Y}}_i \coloneqq \begin{cases} t(\mathbf{Y}_i) \in \mathbb{R}^{k'} & \text{if } Z_i = 1 \\ \emptyset & \text{if } Z_i = 0 \end{cases} \tag{18}$$

for some deterministic function $t : \mathbb{R}^k \to \mathbb{R}^{k'}$. This represents the granular data available for individuals in the cross-section, which may not be the full vector $\mathbf{Y}_i$. In our context, this represents what is observable in the credit card panel, with t being a function which left truncates the individual-level data. $\emptyset$ represents a lack of granular data. Defining the degenerate conditional distribution $\tilde{\mathbf{Y}}_i | Z_i = 0$ where $\tilde{\mathbf{Y}}_i = \emptyset$ with probability 1 ensures that the quantity $\tilde{\mathbf{Y}}_i$ is well-defined even when $Z_i = 0$.

Further define the random variable $\mathbf{D}_N$ as:

$$\mathbf{D}_N \coloneqq \sum_{i=1}^{N} s(\mathbf{Y}_i) \in \mathbb{R}^q \tag{19}$$

for some deterministic function $s : \mathbb{R}^k \to \mathbb{R}^q$. This is the aggregate data, which is a collection of population-level summary statistics of individual-level behaviors. In our empirical application for example, $\mathbf{D}_N$ represents the three customer-related metrics that Spotify discloses in their SEC filings – the total number of customers that are acquired and lost during particular quarters, and the total size of the subscriber base at the end of those quarters.

In our aggregate-disaggregate data fusion setting, the following data are observed:

- Z_i for all $i = 1, 2, \ldots, N$ (whether each population member is in the cross-section)
- $\tilde{\mathbf{Y}}_i$ for all $i = 1, 2, \ldots, N$ (the granular data $t(\mathbf{Y}_i)$ for those with $Z_i = 1$, a degenerate random variable equal to $\emptyset$ for those with $Z_i = 0$)
- $\mathbf{D}_N$ (the population-level aggregate summary statistics)

Define the finite sample and (normalized) limiting mean and variance of $\mathbf{D}_N$ as follows:

$$\begin{aligned} &\boldsymbol{\mu}_N(\boldsymbol{\theta}) \coloneqq E_{\boldsymbol{\theta}}[\mathbf{D}_N | z_{1:N}, \tilde{\mathbf{y}}_{1:N}] = \textstyle\sum_{i=1}^{N} E_{\boldsymbol{\theta}}[s(\mathbf{Y}_i) | z_i, \tilde{\mathbf{y}}_i] \\ &\boldsymbol{\Sigma}_N(\boldsymbol{\theta}) \coloneqq Var_{\boldsymbol{\theta}}[\mathbf{D}_N | z_{1:N}, \tilde{\mathbf{y}}_{1:N}] = \textstyle\sum_{i=1}^{N} Var_{\boldsymbol{\theta}}[s(\mathbf{Y}_i) | z_i, \tilde{\mathbf{y}}_i] \\ &\boldsymbol{\mu}(\boldsymbol{\theta}) \coloneqq E_{\boldsymbol{\theta}}[s(\mathbf{Y}_i)] \\ &\boldsymbol{\Sigma}(\boldsymbol{\theta}) \coloneqq E_{\boldsymbol{\theta}}\left[Var_{\boldsymbol{\theta}}\left[s(\mathbf{Y}_i) | Z_i, \tilde{\mathbf{Y}}_i\right]\right] \end{aligned} \tag{20}$$

3.2. Consistency

First, define the N-normalized proxy likelihood (Equation 10 in the main text) as:

$$Q_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right) \coloneqq \frac{1}{N}\tilde{\ell}_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right) \tag{21}$$

Note that when normalized by N, the third term of the proxy likelihood becomes a quadratic form of sample means:

$$\begin{aligned}
&\frac{1}{N}\left(-\frac{1}{2}\left(\mathbf{d}-\boldsymbol{\mu}_N\left(\boldsymbol{\theta}\right)\right)'\left(\hat{\boldsymbol{\Sigma}}_N\right)^{-1}\left(d-\mu_N\left(\theta\right)\right)\right)\\
=&-\frac{1}{2}\left(\frac{1}{N}\mathbf{d}-\frac{1}{N}\boldsymbol{\mu}_N\left(\boldsymbol{\theta}\right)\right)'\left(\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N\right)^{-1}\left(\frac{1}{N}\mathbf{d}-\frac{1}{N}\boldsymbol{\mu}_N\left(\boldsymbol{\theta}\right)\right)\\
=&-\frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(s(\mathbf{y}_i)-E_{\boldsymbol{\theta}}\left[s(\mathbf{Y}_i)|\,z_i,\tilde{\mathbf{y}}_i\right]\right)\right)'\left(\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left(s(\mathbf{y}_i)-E_{\boldsymbol{\theta}}\left[s(\mathbf{Y}_i)|\,z_i,\tilde{\mathbf{y}}_i\right]\right)\right)
\end{aligned} \tag{22}$$

Accordingly, define the limiting objective function:

$$\begin{aligned}
Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right) \coloneqq\ & E_{\boldsymbol{\theta}_0}\left[\log\left(P_{\boldsymbol{\theta}}\left(Z_i\right)\right)\right]+E_{\boldsymbol{\theta}_0}\left[\log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i|Z_i\right)\right)\right]\\
&-\frac{1}{2}E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)-E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\right]'\hat{\boldsymbol{\Sigma}}^{-1}E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)-E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\right]
\end{aligned} \tag{23}$$

where $\hat{\boldsymbol{\Sigma}}$ is some $q\times q$ positive definite matrix (not necessarily the true covariance matrix $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$). Decompose the function into two terms, $Q^{(a)}$ and $Q^{(b)}$, defined as follows:

$$\begin{aligned}
Q^{(a)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right) &\coloneqq E_{\boldsymbol{\theta}_0}\left[\log\left(P_{\boldsymbol{\theta}}\left(Z_i\right)\right)\right]+E_{\boldsymbol{\theta}_0}\left[\log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i|Z_i\right)\right)\right]\\
Q^{(b)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right) &\coloneqq -\frac{1}{2}E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)-E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\right]'\hat{\boldsymbol{\Sigma}}^{-1}E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)-E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\right]
\end{aligned} \tag{24}$$

such that $Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)=Q^{(a)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)+Q^{(b)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)$.

We assume the following regularity conditions, standard for extremum estimators. All terms are defined according to the Euclidean metric.

1. $\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N\rightarrow_p\hat{\boldsymbol{\Sigma}}$.

(a) That is, the inverse weight matrix $\hat{\boldsymbol{\Sigma}}_N$ can be data-dependent and vary by N so long as, when normalized by N, it converges in probability to a fixed positive definite matrix.

2. Θ is compact, and $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)$ is positive definite for all $\boldsymbol{\theta}\in\Theta$.
3. $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)$ is continuous w.r.t. $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.
4. $P_{\boldsymbol{\theta}}\left(Z_i=z_i\right)$, $P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i=\tilde{\mathbf{y}}_i|Z_i=z_i\right)$, and $E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]$ are each continuous w.r.t. $\boldsymbol{\theta}$ on Θ (and therefore $Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right)$ is continuous w.r.t. $\boldsymbol{\theta}$ on Θ).
5. $Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right)$ is uniquely maximized at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ (identification restriction).

(a) By Jensen's inequality, $Q^{(a)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)$ is maximized at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. Other maximizers $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$ may exist if they imply the same distribution over $\left(\tilde{\mathbf{Y}}_i',Z_i\right)'$.

(b) Since $\hat{\boldsymbol{\Sigma}}$ is positive definite, $Q^{(b)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)$ is a (negative) quadratic form which is maximized when its input $E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)-E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\right]$ is set to 0, i.e. the moment conditions are satisfied. By the law of iterated expectations, the moment conditions are satisfied when $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, so $\boldsymbol{\theta}_0$ is a maximizer of this term. Other maximizers $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$ may exist if they also satisfy the moment conditions exactly.

(c) Define:

$$\begin{aligned}\Theta_a &\coloneqq \left\{ \boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta} \in \arg\max_{\boldsymbol{\theta}\in\Theta} Q^{(a)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right) \right\} \\ \Theta_b &\coloneqq \left\{ \boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta} \in \arg\max_{\boldsymbol{\theta}\in\Theta} Q^{(b)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right) \right\}\end{aligned} \tag{25}$$

We have already shown that $\boldsymbol{\theta}_0 \in \Theta_a$ and $\boldsymbol{\theta}_0 \in \Theta_b$, so $\boldsymbol{\theta}_0 \in \Theta_a \cap \Theta_b$ which implies that $\boldsymbol{\theta}_0$ is a maximizer of $Q\left(\boldsymbol{\theta}|\hat{\Sigma}\right)$. Thus this identification restriction is equivalent to requiring that $(\Theta_a \cap \Theta_b)\backslash\{\boldsymbol{\theta}_0\} = \emptyset$.

(d) Note that this is a weaker restriction than requiring that the model be identified by maximum likelihood on the panel data alone (which would mean $\Theta_a = \{\boldsymbol{\theta}_0\}$) or by generalized method of moments on the summary statistics alone (which would mean $\Theta_b = \{\boldsymbol{\theta}_0\}$); the maximizer for the panel likelihood and the summary statistic moment conditions need not respectively be unique, so long as the intersection of the two sets of maximizers is unique. We discuss in Section 4 of the main text and in Web Appendix 4 the conditions under which this assumption is likely to hold.

6. $\log\left(P_{\boldsymbol{\theta}}\left(Z_i\right)\right) + \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i|Z_i\right)\right)$ and $s(\mathbf{y}_i) - E_{\boldsymbol{\theta}}\left[s(\mathbf{Y}_i)|Z_i = z_i, \tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i\right]$ are uniformly bounded in expectation:

$$\begin{aligned}E_{\boldsymbol{\theta}_0}\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|\log\left(P_{\boldsymbol{\theta}}\left(Z_i\right)\right) + \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i|Z_i\right)\right)\right|\right] &< \infty \\ E_{\boldsymbol{\theta}_0}\left[\sup_{\boldsymbol{\theta}\in\Theta}\left\|s(\mathbf{Y}_i) - E_{\boldsymbol{\theta}}\left[s(\mathbf{Y}_i)|Z_i, \tilde{\mathbf{Y}}_i\right]\right\|\right] &< \infty\end{aligned} \tag{26}$$

By the uniform law of large numbers, the conditions above imply that the objective function converges in probability uniformly:

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|Q_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right) - Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)\right| \to_p 0 \tag{27}$$

Then, by the standard proof of consistency of extremum estimators (Newey and McFadden 1994), $\hat{\boldsymbol{\theta}}_N^{(1)} \to_p \boldsymbol{\theta}_0$. Since $\hat{\boldsymbol{\theta}}_N^{(1)}$ is consistent, by the continuous mapping theorem, the updated covariance matrix $\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N' = \frac{1}{N}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}_N^{(1)}\left(z_{1:N}, \tilde{\mathbf{y}}_{1:N}, \mathbf{d}|\hat{\boldsymbol{\Sigma}}_N\right)\right)$ is a consistent estimator of the true variance of the summary statistics, i.e. $\frac{1}{N}\hat{\boldsymbol{\Sigma}}'_N \to_p \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$, which is by assumption positive definite. Thus, $\frac{1}{N}\hat{\boldsymbol{\Sigma}}'_N$ converges in probability to a positive definite matrix and so by the same argument as for $\hat{\boldsymbol{\theta}}_N^{(1)}$, $\hat{\boldsymbol{\theta}}_N^{(2)} \to_p \boldsymbol{\theta}_0$. That is, both the one-stage estimator $\hat{\boldsymbol{\theta}}_N^{(1)}$ and the two-stage estimator $\hat{\boldsymbol{\theta}}_N^{(2)}$ are consistent. Any further iterations are also consistent by the same argument.

3.3. Asymptotic Normality

To ensure the asymptotic normality of our estimator, we impose the following additional regularity conditions:

7. $\boldsymbol{\theta}_0$ is on the interior of Θ.

8. $P_{\boldsymbol{\theta}}\left(Z_i = z_i\right)$, $P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i|Z_i = z_i\right)$, and $E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i, \tilde{\mathbf{Y}}_i\right]$ are each twice continuously differentiable w.r.t. $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, with bounded second derivatives in this neighborhood.

9. The Hessian matrix $\mathbf{H}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right) \coloneqq \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)$ evaluated at $\boldsymbol{\theta}_0$ is invertible.

Under these assumptions, by the uniform law of large numbers and the continuous mapping theorem, $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right) \rightarrow_p \mathbf{H}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)$ uniformly in a neighborhood of $\boldsymbol{\theta}_0$.

From this, by first order expansion of the gradient around $\boldsymbol{\theta}_0$, denoting the Hessian at $\boldsymbol{\theta}_0$ as $\mathbf{H} \coloneqq \mathbf{H}\left(\boldsymbol{\theta}_0|\hat{\boldsymbol{\Sigma}}\right)$:

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_N^{(1)} - \boldsymbol{\theta}_0\right) \rightarrow_d \mathbf{H}^{-1}\mathbf{G}_{\theta_0}$$

where $\mathbf{G}_{\boldsymbol{\theta}_0}$ is the asymptotic distribution of $\sqrt{N}\,\frac{\partial}{\partial\boldsymbol{\theta}}Q_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ (Newey and McFadden 1994).

To derive $\mathbf{G}_{\boldsymbol{\theta}_0}$, note that we can write the (variance-stabilized) gradient of the objective function as follows:

$$\begin{aligned}\sqrt{N}\frac{\partial}{\partial\boldsymbol{\theta}}Q_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\log\left(P_{\boldsymbol{\theta}}\left(z_i\right)\right) + \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{y}}_i|z_i\right)\right)\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &+ \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|z_i,\tilde{\mathbf{y}}_i\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)\left(\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(s\left(\mathbf{Y}_i\right) - E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)|z_i,\tilde{\mathbf{y}}_i\right]\right)\right) \quad (28)\end{aligned}$$

Since by assumption $\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N \rightarrow_p \hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}$ is invertible, $\left(\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N\right)^{-1} \rightarrow_p \hat{\boldsymbol{\Sigma}}^{-1}$ by the continuous mapping theorem. Denoting the sample and expected Jacobian of expected summary statistics as follows:

$$\begin{aligned}\mathbf{J}_N &\coloneqq \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|z_i,\tilde{\mathbf{y}}_i\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \mathbf{J} &\coloneqq E_{\boldsymbol{\theta}_0}\left[\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]\end{aligned} \quad (29)$$

By the weak law of large numbers, $\mathbf{J}_N \rightarrow_p \mathbf{J}$. Lastly, denote the sample and expected gradients of the granular data likelihood and the sample and expectated deviations from the conditional expectation of $s\left(\mathbf{Y}_i\right)$ as follows:

$$\begin{aligned}\mathbf{S}_N &\coloneqq \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\log\left(P_{\boldsymbol{\theta}}\left(z_i\right)\right) + \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{y}}_i|z_i\right)\right)\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \mathbf{S} &\coloneqq E_{\boldsymbol{\theta}_0}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\log\left(P_{\boldsymbol{\theta}}\left(z_i\right)\right) + \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{y}}_i|z_i\right)\right)\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \\ \mathbf{M}_N &\coloneqq \frac{1}{N}\sum_{i=1}^{N}\left(s\left(\mathbf{Y}_i\right) - E_{\boldsymbol{\theta}}\left[s\left(\mathbf{Y}_i\right)|z_i,\tilde{\mathbf{y}}_i\right]\right) \\ \mathbf{M} &\coloneqq E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right) - E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\right]\end{aligned} \quad (30)$$

As discussed in Assumption 4, $\boldsymbol{\theta}_0$ is a maximizer of $Q^{(a)}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}\right)$, and so by the differentiability of this term, the first order condition must be satisfied at $\boldsymbol{\theta}_0$, and so by the dominated convergence theorem, $\mathbf{S} = 0$. Additionally, from the law of iterated expectations, it is easy to see that $\mathbf{M} = 0$. Thus $\mathbf{S}_N$ and $\mathbf{M}_N$ are both mean zero, and so by the multivariate central limit theorem:

$$\sqrt{N}\begin{pmatrix}\mathbf{S}_N \\ \mathbf{M}_N\end{pmatrix} \rightarrow_d \mathcal{N}\left(\mathbf{0},\boldsymbol{\Omega}\right) \quad (31)$$

where

$$\boldsymbol{\Omega} \coloneqq Var_{\boldsymbol{\theta}_0}\begin{pmatrix}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\log\left(P_{\boldsymbol{\theta}}\left(z_i\right)\right) + \log\left(P_{\boldsymbol{\theta}}\left(\tilde{\mathbf{y}}_i|z_i\right)\right)\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ s\left(\mathbf{Y}_i\right) - E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)|Z_i,\tilde{\mathbf{Y}}_i\right]\end{pmatrix} \quad (32)$$

Then, from the continuous mapping theorem:

$$\sqrt{N}\frac{\partial}{\partial\boldsymbol{\theta}}Q_N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\Sigma}}_N\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \sqrt{N}\mathbf{S}_N + \mathbf{J}_N\left(\frac{1}{N}\hat{\boldsymbol{\Sigma}}_N\right)^{-1}\left(\sqrt{N}\mathbf{M}_N\right) \rightarrow_d \mathcal{N}\left(\mathbf{0},\mathbf{T}\boldsymbol{\Omega}\mathbf{T}'\right) \quad (33)$$

where $\mathbf{T} \coloneqq \left(\mathbb{I}_{p\times p}, \mathbf{J}\hat{\mathbf{\Sigma}}^{-1}\right)$ is the linear transformation matrix that maps the limiting distribution of $\sqrt{N}\left(\mathbf{S}'_N, \mathbf{M}'_N\right)'$ to the limiting distribution of $\sqrt{N}\frac{\partial}{\partial \boldsymbol{\theta}} Q_N\left(\boldsymbol{\theta}|\hat{\mathbf{\Sigma}}_N\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.

Putting all of this together:

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_N^{(1)} - \boldsymbol{\theta}_0\right) \rightarrow_d \mathcal{N}\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{T}\mathbf{\Omega}\mathbf{T}'\mathbf{H}^{-1}\right) \tag{34}$$

Thus, $\hat{\boldsymbol{\theta}}_N^{(1)}$ converges to $\boldsymbol{\theta}_0$ at rate $O_p\left(N^{-1/2}\right)$ and achieves asymptotic normality. This result also holds for $\hat{\boldsymbol{\theta}}_N^{(2)}$, since as noted before $\frac{1}{N}\hat{\mathbf{\Sigma}}'_N \rightarrow_p \mathbf{\Sigma}\left(\boldsymbol{\theta}_0\right)$, which is positive definite by assumption. Any additional iterations (e.g. updating $\hat{\mathbf{\Sigma}}_N$ based on $\hat{\boldsymbol{\theta}}_N^{(2)}$ and estimating a third time) will yield the same asymptotic distribution as $\hat{\boldsymbol{\theta}}_N^{(2)}$.

Unlike for the generalized method of moments, because of possible correlations between the gradient of the granular data log-likelihood and the deviation from the conditional expectation of $s\left(\mathbf{Y}_i\right)$, it is difficult to establish a general result on the optimal choice of weight matrix $\hat{\mathbf{\Sigma}}^{-1}$. Intuitively, however, the use of a consistent estimator of the true precision $\mathbf{\Sigma}\left(\boldsymbol{\theta}_0\right)^{-1}$ as employed by $\hat{\boldsymbol{\theta}}_N^{(2)}$ should perform well so long as these correlations are not too large, as it establishes the "correct" (based on the normal likelihood approximation motivation) relative weight of the summary statistics and granular data, and mirrors the optimal choice of weight matrix in the generalized method of moments (Hansen 1982). Empirically in our parameter recovery simulation studies we find that the two-stage estimator achieves on average a 7.2% parameter-wise reduction in absolute estimation error (MAE) compared to the one-stage estimator. This indicates that using a two-stage estimator may improve empirical performance compared to a naive initialization of the weight matrix. For our baseline set of parameters in the parameter recovery simulation study, we found that continuing estimation for a third stage of optimization resulted in no change in parameter estimation error within 3 significant figures.

To compute asymptotically valid standard errors, we need consistent estimators of $\mathbf{H}$, $\mathbf{T}$, and $\mathbf{\Omega}$. $\mathbf{H}$ and $\mathbf{T}$ can be approximated by their sample analogues: the Hessian of $Q_N\left(\hat{\boldsymbol{\theta}}_N^{(1)}|\hat{\mathbf{\Sigma}}_N\right)$ for $\mathbf{H}$ and $\left(\mathbb{I}_{p\times p}, \frac{1}{N}\mathbf{J}_N\hat{\mathbf{\Sigma}}_N\right)$ for $\mathbf{T}$ in the case of $\hat{\boldsymbol{\theta}}_N^{(1)}$, and $Q_N\left(\hat{\boldsymbol{\theta}}_N^{(2)}|\mathbf{\Sigma}\left(\hat{\boldsymbol{\theta}}_N^{(2)}\right)\right)$ and $\left(\mathbb{I}_{p\times p}, \mathbf{J}_N\mathbf{\Sigma}\left(\hat{\boldsymbol{\theta}}_N^{(2)}\right)\right)$ in the case of $\hat{\boldsymbol{\theta}}_N^{(2)}$. However, there is no sample analogue of $\mathbf{\Omega}$, because by definition $s\left(\mathbf{Y}_i\right)$ is not observed. Instead, we can approximate it through simulation: we can simulate uncensored outcomes $\left(\mathbf{Y}'_i, Z_i\right)'$ from $F_{\boldsymbol{\theta}}$ at the estimated value of $\boldsymbol{\theta}$ and compute the empirical covariance matrix $\hat{\mathbf{\Omega}}$ from the simulated datapoints. This gives asymptotically correct standard errors when the number of simulations used to compute $\hat{\mathbf{\Omega}}$ is large.

Web Appendix 4: Semiparametric Identification of Selection Function

In our empirical specification, we use a logit-linear specification for the selection function $P(Z_i = 1|\boldsymbol{\lambda}_i)$. However, we argued in Section 4.3 that more general selection functions can also be identified with sufficiently rich data. Here, we discuss some formal conditions under which general selection functions are identified for our model.

In our specification, we started with a specification of the population heterogeneity distribution $g(\boldsymbol{\lambda}_i)$, then computed the panel heterogeneity distribution $g(\boldsymbol{\lambda}_i|Z_i = 1)$ by re-weighting the population distribution via

Bayes theorem. However, we can equivalently reparameterize our model in terms of the panel heterogeneity distribution $g(\boldsymbol{\lambda}_i|Z_i=1)$ and overall size of panel $P(Z_i=1)$. To see this, note that by Bayes theorem:

$$g(\boldsymbol{\lambda}_i|Z_i=1)=\frac{P(Z_i=1|\boldsymbol{\lambda}_i)g(\boldsymbol{\lambda}_i)}{P(Z_i=1)}.$$

Then as long as $P(Z_i=1|\boldsymbol{\lambda}_i)>0$ for all $\boldsymbol{\lambda}_i$ on the support of $g(\boldsymbol{\lambda}_i)$, the density of the population heterogeneity distribution $g(\boldsymbol{\lambda}_i)$ is a re-weighting of the panel distribution (weighted by the inverse selection probability):

$$g(\boldsymbol{\lambda}_i)=g(\boldsymbol{\lambda}_i|Z_i=1)\cdot\left(\frac{P(Z_i=1)}{P(Z_i=1|\boldsymbol{\lambda}_i)}\right).$$

Consider a general model of the panel heterogeneity distribution $g_{\boldsymbol{\varphi}}(\boldsymbol{\lambda}_i|Z_i=1)$ parameterized by vector $\boldsymbol{\varphi}\in\boldsymbol{\Phi}$ and a general model of the selection function $P_{\boldsymbol{\psi}}(Z_i=1|\boldsymbol{\lambda}_i)$ parameterized by vector $\boldsymbol{\psi}\in\boldsymbol{\Psi}$. Let $\boldsymbol{\eta}\in\mathcal{H}$ be the vector of homogeneous parameters, and let $\pi^{(Z)}\coloneqq P(Z_i=1)\in(0,1)$ be the proportion of the population in the panel, such that the full vector of model parameters θ is the concatenation of $\boldsymbol{\eta}$, $\pi^{(Z)}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\psi}$.

Under this reparameterization, we can construct a simple two-stage estimator, which first estimates $\boldsymbol{\eta}$, $\boldsymbol{\varphi}$, and $\pi^{(Z)}$ on the panel data, then estimates $\boldsymbol{\psi}$ on the aggregate data, plugging in the estimates $\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{\varphi}}$, and $\hat{\pi}^{(Z)}$ from the first stage. Note that this is not actually the estimation method that we propose and use in our applications; rather, it is a convenient expositional tool for understanding which parameters are identified by which parts of the data. Our proposed method utilizes both the aggregate data and panel data jointly for estimation, allowing for better statistical efficiency than this convenient expositional estimator.

Assume the same regularity conditions as in Web Appendix 3.2, except for the identification restriction (condition 5), but in terms of the likelihood function of the reparameterized model above. Additionally assume the following:

1. $E_{\boldsymbol{\eta},\boldsymbol{\varphi}}\left[\log\left(P_{\boldsymbol{\eta},\boldsymbol{\varphi}}\left(\tilde{\mathbf{Y}}_i|Z_i\right)\right)\right]$ is uniquely maximized at $\boldsymbol{\eta}=\boldsymbol{\eta}_0,\boldsymbol{\varphi}=\boldsymbol{\varphi}_0$, where $\boldsymbol{\eta}_0$ and $\boldsymbol{\varphi}_0$ are the true values of $\boldsymbol{\eta}$ and $\boldsymbol{\varphi}$, respectively (partial identification of panel-specific parameters).

2. $P_{\boldsymbol{\psi}}(Z_i=1|\boldsymbol{\lambda}_i)>0$ at every point on the support of $\boldsymbol{\lambda}_i$ in the aggregate population, for all $\boldsymbol{\psi}\in\boldsymbol{\Psi}$ (overlap condition).

The first condition states that the panel data is rich enough to identify the homogeneous parameters $\boldsymbol{\eta}$ and the panel heterogeneity distribution parameters $\boldsymbol{\varphi}$. As we argue in Section 4.3, this should hold as long as the panel data covers a long enough time scale, since long-term trends in acquisition and churn patterns observed in the panel data allow for identification of the homogeneous acquisition/churn process characteristics, as well as the panel-specific heterogeneity distribution. Intuitively, the second condition states that all values of $\boldsymbol{\lambda}_i$ that may appear in the aggregate population distribution are represented in the panel data: that is, there are no segments of the population distribution of $\boldsymbol{\lambda}_i$ which have 0 probability of being selected into the panel. This type of assumption is necessary to avoid divide-by-zero issues.

Under these conditions, the maximum likelihood estimator

$$\{\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\varphi}}\}=\underset{\{\boldsymbol{\eta},\boldsymbol{\varphi}\}\in\mathcal{H}\times\boldsymbol{\Phi}}{\arg\max}\sum_{\{i|Z_i=1\}}\log\left(P_{\boldsymbol{\eta},\boldsymbol{\varphi}}\left(\tilde{\mathbf{Y}}_i|Z_i\right)\right)$$

is a consistent estimator of $\{\boldsymbol{\eta}_0,\boldsymbol{\varphi}_0\}$, by the standard proof of consistency of extremum estimators (Newey and McFadden 1994).

Next, $\pi^{(Z)}$ is simply the probability parameter of a Bernoulli distribution, and so its maximum likelihood estimate is the sample proportion of panel members relative to the population:

$$\hat{\pi}^{(Z)} = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

which is consistent. While the summation is technically over all population members, this estimate only requires knowing the overall size of the population (which in our application is assumed known) and the size of the panel, without needing any population-level data. Thus, maximum likelihood on the panel data gives consistent estimates of $\boldsymbol{\eta}_0$, $\boldsymbol{\varphi}_0$, and $\pi_0^{(Z)}$.

This leaves the second-stage estimation of the selection parameters $\boldsymbol{\psi}$. As shown above, the population heterogeneity distribution can be written as a reweighting of the panel heterogeneity distribution, such that the aggregate moments can be written in terms of the selection function and the above parameters without requiring the population heterogeneity distribution directly:

$$E\left[s\left(\mathbf{Y}_i\right)\right] = \int_{\mathbb{R}_+^4} E\left[s\left(\mathbf{Y}_i\right)|\boldsymbol{\lambda}_i\right] g\left(\boldsymbol{\lambda}_i\right) d\boldsymbol{\lambda}_i = \int_{\mathbb{R}_+^4} E\left[s\left(\mathbf{Y}_i\right)|\boldsymbol{\lambda}_i\right] g(\boldsymbol{\lambda}_i|Z_i=1)\left(\frac{P(Z_i=1)}{P(Z_i=1|\boldsymbol{\lambda}_i)}\right) d\boldsymbol{\lambda}_i$$

As discussed in Web Appendix 2, we can efficiently compute the inner expectation conditional on $\boldsymbol{\lambda}_i$ efficiently via a recursive algorithm, and the integral over the density on the right-hand side is easily approximated numerically using a minor modification of the Halton sequence procedure described in Web Appendix 2.3; thus, this equation is feasible to compute in estimation. As a result, we can plug in the first stage estimates to compute aggregate moments as a function of selection parameters $\boldsymbol{\psi}$:

$$E_{\boldsymbol{\psi},\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\varphi}},\hat{\pi}^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right] = \int_{\mathbb{R}_+^4} E_{\hat{\boldsymbol{\eta}}}\left[s\left(\mathbf{Y}_i\right)|\boldsymbol{\lambda}_i\right] g_{\hat{\boldsymbol{\varphi}}}(\boldsymbol{\lambda}_i|Z_i=1)\left(\frac{\hat{\pi}^{(Z)}}{P_{\boldsymbol{\psi}}(Z_i=1|\boldsymbol{\lambda}_i)}\right) d\boldsymbol{\lambda}_i$$

Our aggregate data consist of sample moments $\mathbf{D}_N = \sum_{i=1}^{N} s\left(\mathbf{Y}_i\right)$, such that $\boldsymbol{\psi}$ can be estimated by the generalized method of moments (GMM):

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\arg\min}\left(\frac{1}{N}\mathbf{D}_N - E_{\boldsymbol{\psi},\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\varphi}},\hat{\pi}^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]\right)' W_N \left(\frac{1}{N}\mathbf{D}_N - E_{\boldsymbol{\psi},\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\varphi}},\hat{\pi}^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]\right)$$

where W_N is a positive definite weight matrix which asymptotically converges in probability to a fixed positive definite matrix $W_N \to_p W$. Note that the selection parameter vector $\boldsymbol{\psi}$ only enters the equation for $E_{\boldsymbol{\psi},\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\varphi}},\hat{\pi}^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]$ in the inverse probability weight, emphasizing the intuition for how the selection function is identified: taking the moments implied by the panel data, and finding the appropriate reweighting that eliminates the discrepancies between the panel data and aggregate moments.

Since the first-stage estimation is consistent, $\hat{\boldsymbol{\eta}} \to_p \boldsymbol{\eta}_0$, $\hat{\boldsymbol{\varphi}} \to_p \boldsymbol{\varphi}_0$, and $\hat{\pi}^{(Z)} \to_p \pi_0^{(Z)}$, and by assumption the aggregate moments are continuous w.r.t. the model parameters, such that by the continuous mapping theorem, $\hat{\boldsymbol{\psi}}$ is also consistent as long as the moment conditions imposed by $E_{\boldsymbol{\psi},\boldsymbol{\eta}_0,\boldsymbol{\varphi}_0,\pi_0^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]$ satisfy the usual identification conditions of GMM (Hansen 1982).

In particular, defining the limiting GMM objective function:

$$J\left(\boldsymbol{\psi}\right) = \left(E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)\right] - E_{\boldsymbol{\psi},\boldsymbol{\eta}_0,\boldsymbol{\varphi}_0,\pi_0^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]\right)' W \left(E_{\boldsymbol{\theta}_0}\left[s\left(\mathbf{Y}_i\right)\right] - E_{\boldsymbol{\psi},\boldsymbol{\eta}_0,\boldsymbol{\varphi}_0,\pi_0^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]\right)$$

By construction, J is non-negative, and $J(\boldsymbol{\psi}_0) = 0$, and so $\boldsymbol{\psi}_0$ is a minimizer of J; thus, the identification conditions are satisfied if and only if $J(\boldsymbol{\psi}_0) > 0$ for all $\boldsymbol{\psi} \neq \boldsymbol{\psi}_0$ in $\boldsymbol{\Psi}$ (i.e. no other configuration of selection

parameters can exactly explain the discrepancies between the panel data and the aggregate moments). Under this condition, $\hat{\boldsymbol{\psi}} \rightarrow_p \boldsymbol{\psi}_0$, and so the two-stage estimator is consistent with $\hat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$, implying that the model is identified.

This raises the question of when $\boldsymbol{\psi}_0$ will be a unique minimizer. At minimum, there must be at least as many moment conditions as there are selection parameters, i.e. the dimension of s must be at least as large as the dimension of $\boldsymbol{\psi}$, and the Jacobian $\nabla_{\boldsymbol{\psi}} E_{\boldsymbol{\psi},\boldsymbol{\eta}_0,\boldsymbol{\varphi}_0,\pi_0^{(Z)}}\left[s\left(\mathbf{Y}_i\right)\right]$ evaluated at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ must have full column rank. This ensures that $\boldsymbol{\psi}$ is *locally* identified, in the sense that in some neighborhood of $\boldsymbol{\psi}_0$, no other value of $\boldsymbol{\psi}$ satisfies the moment conditions exactly (Newey and McFadden 1994).

Beyond these sufficient conditions for local identification, it is difficult to establish guarantees about global identification for general selection functions, but intuitively we require that each parameter in $\boldsymbol{\psi}$ results in a different panel vs. population discrepancy that manifests in the observable aggregate data. For instance, as discussed in Section 4.3, our proposed model specification with first-order selection biases implies level shifts between the acquisition and retention curves in the panel vs. aggregate data. Since initial vs. repeat customers are not separated out in the aggregate summaries ADD_q and $LOSS_q$, we need to observe acquisition and retention counts at different points in time as the composition of customers changes from more initial to more repeat customers, so as to separate out how much of the level shifts in acquisitions and churns are attributable to selection bias in initial vs. repeat behavioral propensities. Thus, in principle, our model could be identified with as few as 4 aggregate moments to supplement the panel data: 2 observations each of ADD_q and $LOSS_q$ that are temporally separated.

Model specifications that allow for second-order selection biases would require more aggregate moments that demonstrate sequential correlations between acquisition and churn counts, since as described in Section 4.3, observing such patterns is necessary for identifying variance and other higher-order characteristics of the heterogeneity distributions based on the aggregate data. If there are enough moments to reflect higher-order heterogeneity patterns in the aggregate data, then these moments allow for identification of the model, since the selection function can be identified off of the discrepancies between the higher-order heterogeneity patterns in the panel vs. aggregate data.

Web Appendix 5: Detailed Simulation Study Results

In this web appendix, we give further details of our simulation study implementation and results. Web Appendix 5.1 reports the parameter settings used for the predictive simulation study and gives further description of our implementation. Web Appendix 5.2 reports the compute times of the proxy likelihood objective function compared to those of benchmark methods. Web Appendices 5.3 and 5.4 provide further detail on the results of our predictive simulation study and parameter recovery simulation study, respectively.

5.1. Implementation Details

Predictive Simulation Parameter Settings

The eight parameter settings that we vary through a full factorial design are as follows:

1. Initial acquisition baseline: We consider $(\lambda_0^{(IA)}, c^{(IA)})$ values of (.001, 1.2) and (.1,.8).
2. Initial churn baseline: We consider $(\lambda_0^{(IC)}, c^{(IC)})$ values of (.1,.8) and (.2,.6).

3. Repeat acquisition baseline: We consider $(\lambda_0^{(RA)}, c^{(RA)})$ values of (.1,.8) and (.2,.6).
4. Repeat churn baseline: We consider $(\lambda_0^{(RC)}, c^{(RC)})$ values of (.1,.8) and (.2,.6).
5. Initial process variance: We consider values of 1 and 2 for $(\sqrt{\sigma^{2(IA)}}, \sqrt{\sigma^{2(IC)}})$
6. Repeat process variance: We consider values of 1 and 2 for $(\sqrt{\sigma^{2(RA)}}, \sqrt{\sigma^{2(RC)}})$
7. Within-process correlation: We consider values of 0% and 20% for all within-process correlations (i.e., $\sigma^{(A,B)}/(\sqrt{\sigma^{2(A)}}\sqrt{\sigma^{2(B)}})$ for $(A,B) = (IA, RA)$ and (IC, RC).
8. Across-process correlation: We consider values of 0% and 20% for all across-process correlations (i.e., $\sigma^{(A,B)}/(\sqrt{\sigma^{2(A)}}\sqrt{\sigma^{2(B)}})$ for $(A,B) = (IA, IC)$, (IA, RC), (IC, RA), and (RA, RC).

Panel selection bias was varied through $\boldsymbol{\beta}^{(Z)}$. No panel bias is equivalent to setting $\boldsymbol{\beta}^{(Z)} = (0,0,0,0)$. We equated moderate and severe panel bias to $\boldsymbol{\beta}^{(Z)} = (1/2, 1/2, -1/2, -1/2)$ and $(1,1,-1,-1)$, respectively. We fix both $\pi^{(IA)}$ and $\pi^{(RA)}$ at 90%.

Estimation Details

We estimate the parameters using the PAN method by maximum likelihood. We then take the implied customer base projections for the panel members and gross them up to the size of the target population. For both AGG and MPL, we employ the two-step estimator described in Section 3.2 (for AGG, this reduces to two-stage efficient Generalized Method of Moments). We initialize the covariance matrix $\hat{\boldsymbol{\Sigma}}_N^{(1)}$ to $.25 \times N$ multiplied by an identity matrix whose dimension is equal to the number of observed aggregate summary statistics, as this is equal to the maximum variance of the sum of N Bernoulli random variables. We obtain initial parameter estimates for the two methods. We use these parameters to initialize the second-stage covariance matrix which is used to obtain our final parameter estimate $\tilde{\boldsymbol{\theta}}$. While we specified an upper limit on the number of iterations that the algorithm would run for before forcing convergence at 2,000, this limit was only reached 4 times (0.04% of the estimations).

Simulating from $g(\boldsymbol{\lambda})$

It is well known that simulation-based integration methods can yield biased estimators when a finite number of simulations is used (Train 2009). Because our objective is to study the finite sample performance of the MPL method and not the performance of simulation-based integration methods, for our parameter recovery simulation study we simulate individual $\boldsymbol{\lambda}$s by resampling from the same Halton draws that we use to compute the proxy likelihood, rather than sampling from the continuous lognormal distribution. This can be interpreted as a correctly specified integration procedure where the true heterogeneity distribution $g(\boldsymbol{\lambda})$ is a discrete grid of K support points encoded by the first K terms of the Halton sequence, passed through an affine transformation parameterized by $\boldsymbol{\lambda}_0$ and $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}$. We obtain similar results using a continuous lognormal distribution as the data generating process when we use a sufficiently large K. In benchmarking predictive accuracy, bias in the parameter estimates is irrelevant as long as the resulting predictions are still accurate, so for the predictive simulation study we take $g(\boldsymbol{\lambda})$ to be a continuous lognormal distribution.

Table 1 Simulation study: Median objective function evaluation time (seconds) by method and data setting

Setting	Value	PAN	AGG	MPL
Overall		0.617	0.090	0.639
M	36	0.450	0.037	0.467
	60	0.658	0.089	0.673
	84	0.760	0.165	0.784
	108	0.915	0.289	0.890
Panel %	1.0%	0.193	0.089	0.202
	5.0%	0.662	0.090	0.681
	10.0%	1.057	0.089	1.068

Setting	Value	PAN	AGG	MPL
N	20K	0.190	0.090	0.201
	100K	0.571	0.090	0.593
	500K	2.640	0.090	2.572
	2,500K	12.772	0.089	12.599
Panel Bias	High	0.609	0.089	0.642
	Medium	0.662	0.090	0.680
	None	0.436	0.090	0.576

5.2. Comparison of Compute Times

We provide the median objective function compute times for each method used in the predictive simulation study, averaging across all data settings as well as for each of the data settings individually in Table 1. These results average across all 256 parameter settings associated with each data setting. All computing time figures were obtained using an Intel Xeon processor with a 3.00GHz clock speed.

As noted previously, these results suggest that MPL's compute time is only marginally higher than that of PAN, while AGG's compute time is materially lower. Compute time increases quadratically with M, as expected theoretically. It increases monotonically with N and panel proportion because the MPL method evaluates the likelihood of the panel data, the size of which is a function of N.

5.3. Predictive Simulation Study Results

In Tables 2, 3, 4, and 5, we provide MAPE figures by method and aggregated summary statistic (hereafter referred to interchangeably as "disclosure") for each of the four data settings: panel selection bias, M, panel size as a percentage of N, and N.

5.4. Parameter Recovery Simulation Study Results

In this section, we provide detailed statistics summarizing the parameter recovery performance of the MPL method. We provide an additional two sets of results:

- The first contains all parameter recovery figures on a parameter-by-parameter basis for five different sets of parameters – the baseline setting studied in Section 5.2, and an additional four randomly selected parameter settings from the large-scale simulation analysis in Section 5. For each parameter within each of the resulting five parameter settings, we provide the true parameter value ("True Param." in the tables that follow), the MPL method's estimate of the true parameter values ("Est Param."), bias ("Bias"), and the standard deviation of the parameter estimates ("SD of Ests"). Results for the baseline parameter setting are in Table 6 while the analogous results for parameter settings 2 through 5 are in Tables 7, 8, 9, and 10, respectively.
- The second extends the results shown in Table 3 of Section 5.2, providing the parameter recovery figures for the MPL method when the selection model is misspecified. Recall that our proposed model assumes that the logit probability of being selected into the panel is a linear function of the log baseline propensities to initial and repeat acquire and churn (Equation 4 in the main body of the paper). In Table 11, we provide the

Table 2 MAPE by method and disclosure, varying selection bias, averaging across parameter settings

Disc.	Panel Bias	PAN	AGG	MPL
QIA	High	42.6%	23.0%	10.6%
	Medium	32.0%	23.3%	8.9%
	None	10.6%	26.0%	10.3%
QIC	High	13.4%	24.0%	8.8%
	Medium	10.9%	22.6%	8.0%
	None	5.8%	25.6%	7.1%
QRA	High	24.1%	9.6%	4.2%
	Medium	17.4%	8.9%	3.9%
	None	7.4%	10.2%	3.2%
QRC	High	26.4%	10.7%	4.3%
	Medium	17.2%	10.1%	4.3%
	None	8.0%	11.8%	3.6%
ADD	High	26.4%	2.4%	2.1%
	Medium	18.2%	2.4%	2.0%
	None	5.5%	2.3%	1.8%
$LOSS$	High	21.5%	2.6%	2.1%
	Medium	14.6%	2.6%	2.1%
	None	6.2%	2.5%	1.9%
END	High	129.9%	0.7%	0.5%
	Medium	88.7%	1.0%	0.5%
	None	2.5%	0.7%	0.5%

Table 3 MAPE by method and disclosure, varying M, averaging across parameter settings

Disc.	M	PAN	AGG	MPL
QIA	36	34.4%	22.2%	10.7%
	60	30.5%	22.9%	9.2%
	84	31.4%	25.0%	8.1%
	108	31.5%	28.2%	8.2%
QIC	36	12.7%	24.7%	9.7%
	60	10.8%	22.9%	8.0%
	84	9.4%	22.1%	7.0%
	108	8.8%	23.3%	6.7%
QRA	36	20.0%	13.2%	6.1%
	60	17.5%	9.1%	3.9%
	84	14.7%	7.5%	3.0%
	108	13.9%	6.1%	2.7%
QRC	36	19.7%	17.2%	7.2%
	60	17.5%	10.2%	4.1%
	84	15.3%	8.1%	3.2%
	108	14.1%	6.5%	2.8%
ADD	36	22.1%	3.8%	2.3%
	60	18.0%	2.3%	2.0%
	84	15.7%	2.3%	2.0%
	108	14.0%	1.8%	2.0%
$LOSS$	36	15.9%	4.3%	2.4%
	60	14.7%	2.5%	2.0%
	84	13.2%	2.4%	2.0%
	108	12.1%	2.0%	2.0%
END	36	89.1%	1.5%	0.8%
	60	83.4%	0.9%	0.5%
	84	87.7%	0.8%	0.4%
	108	86.6%	0.4%	0.4%

bias and variance of this estimator when the true logit-transformed panel selection probability is a linear, square root or quadratic function of $\log(\boldsymbol{\lambda})$, but we estimate the model assuming the linear specification.[2]

All results are averaged across 30 replicates simulated from each set of parameter values. We do not include the results for the method trained only upon the aggregate data, because in all parameter settings, the bias and variance of this estimator was at least one order of magnitude worse than the MPL method. As was the case in Section 5.2, we also did not include the results for the method trained only upon the panel data because this method is asymptotically inconsistent in the presence of any selection bias.

The results from Tables 6 through 10 indicate that the main findings noted in Section 5.2 – low finite sample bias and reasonable empirical identification (as measured by low sample-to-sample variation in estimated parameter values) – largely hold on a parameter-by-parameter basis, and are robust to the specific parameter values we use to generate the data.

The results from Table 11 show that when the true selection mechanism is linear, bias and variance are generally small, and empirical coverage is near its theoretical target level. When the true selection mechanism

[2] $\log(\boldsymbol{\lambda})$ can take on negative values, so the true selection function has as an input $\text{sign}\left(\log\left(\boldsymbol{\lambda}_i\right)\right) \times \sqrt{\left|\log\left(\boldsymbol{\lambda}_i\right)\right|}$.

Table 4 MAPE by method and disclosure, varying panel size as a percentage of N, averaging across parameter settings

Disc.	Panel %	PAN	AGG	MPL
QIA	1.0%	35.9%	25.6%	10.9%
	5.0%	30.5%	23.3%	9.2%
	10.0%	30.2%	23.2%	7.2%
QIC	1.0%	13.4%	25.1%	10.8%
	5.0%	10.3%	22.7%	7.9%
	10.0%	10.7%	23.3%	6.0%
QRA	1.0%	21.7%	10.3%	4.7%
	5.0%	16.6%	8.9%	3.8%
	10.0%	17.3%	9.2%	3.6%
QRC	1.0%	22.7%	11.2%	5.4%
	5.0%	16.8%	10.2%	4.1%
	10.0%	15.2%	10.2%	3.7%
ADD	1.0%	21.7%	2.3%	1.9%
	5.0%	17.4%	2.4%	2.0%
	10.0%	17.6%	2.4%	1.9%
$LOSS$	1.0%	18.8%	2.5%	2.0%
	5.0%	14.1%	2.6%	2.1%
	10.0%	13.7%	2.6%	2.0%
END	1.0%	106.7%	0.7%	0.5%
	5.0%	82.9%	1.0%	0.5%
	10.0%	77.9%	0.7%	0.5%

Table 5 MAPE by method and disclosure, varying N, averaging across parameter settings

Disc.	N	PAN	AGG	MPL
QIA	20K	31.5%	26.8%	13.7%
	100K	31.2%	24.8%	9.3%
	500K	30.2%	18.7%	6.8%
	2,500K	29.8%	15.2%	5.8%
QIC	20K	13.1%	29.9%	13.5%
	100K	10.7%	24.1%	8.0%
	500K	9.0%	17.9%	5.4%
	2,500K	9.7%	13.1%	4.9%
QRA	20K	17.7%	11.1%	6.4%
	100K	17.2%	9.5%	3.9%
	500K	15.6%	7.0%	2.6%
	2,500K	17.2%	5.3%	2.1%
QRC	20K	19.8%	13.9%	7.4%
	100K	17.4%	10.9%	4.3%
	500K	14.9%	7.5%	2.5%
	2,500K	15.5%	5.3%	2.0%
ADD	20K	17.9%	5.0%	4.0%
	100K	17.8%	2.5%	2.0%
	500K	17.1%	1.1%	1.1%
	2,500K	18.0%	0.5%	0.9%
$LOSS$	20K	16.1%	5.2%	4.1%
	100K	14.6%	2.7%	2.1%
	500K	12.7%	1.3%	1.1%
	2,500K	13.7%	0.6%	0.9%
END	20K	86.4%	3.5%	1.0%
	100K	83.7%	0.8%	0.5%
	500K	87.1%	0.3%	0.3%
	2,500K	88.2%	0.2%	0.2%

is a square root or quadratic function of $\log(\boldsymbol{\lambda})$ but our estimator assumes that this relationship is linear, bias increases, variance remains approximately the same, and coverage degrades. For example, the mean absolute bias of the mean heterogeneity and variance heterogeneity parameters are roughly 1% when the selection function is correctly specified, moving up to roughly 10% under both misspecified selection specifications. Coverage somewhat degrades when the true selection function is a square root function of $\log(\boldsymbol{\lambda})$, falling to roughly 90% from the target level of 95%, but falls much more significantly when the true selection is a quadratic function of $\log(\boldsymbol{\lambda})$, falling to roughly 65%.

While parameter recovery performance weakens under misspecification, these results nevertheless suggest that the overall performance of the MPL estimator is relatively robust to the exact specification of the selection function within reason (e.g., if we can assume the relationship between the selection probability and λ is monotone). Should the selection function be more severely misspecified, bias and coverage weakens more significantly. Modelers should thoroughly validate their proposed model, as we demonstrate in Section 6 of the paper, to guard themselves against the possibility of weak empirical performance due to model misspecification.

Table 6 Parameter recovery analysis: proposed method, all parameters, baseline scenario

Parameter	True Param.	Est Param.	Bias	SD of Ests
$\lambda_0^{(IA)}$	0.0505	0.0519	0.001	0.007
$c^{(IA)}$	1.0000	0.9949	−0.005	0.044
$\pi^{(IA)}$	0.9000	0.8992	−0.001	0.009
$\lambda_0^{(RA)}$	0.1500	0.1509	0.001	0.014
$c^{(RA)}$	0.7000	0.7013	0.001	0.021
$\pi^{(RA)}$	0.9000	0.9007	0.001	0.009
$\lambda_0^{(IC)}$	0.1500	0.1509	0.001	0.015
$c^{(IC)}$	0.7000	0.7022	0.002	0.042
$\lambda_0^{(RC)}$	0.1500	0.1503	0.000	0.016
$c^{(RC)}$	0.7000	0.6984	−0.002	0.026
$\sigma_{\lambda^{(IA)}}$	1.5000	1.4772	−0.023	0.125
$\rho_\lambda^{(IA,RA)}$	0.1000	0.0964	−0.004	0.039
$\sigma_{\lambda^{(RA)}}$	1.5000	1.5085	0.008	0.065
$\rho_\lambda^{(IA,IC)}$	0.1000	0.0968	−0.003	0.032
$\rho_\lambda^{(RA,IC)}$	0.1000	0.0821	−0.018	0.035
$\sigma_{\lambda^{(IC)}}$	1.5000	1.4945	−0.005	0.147
$\rho_\lambda^{(IA,RC)}$	0.1000	0.1029	0.003	0.040
$\rho_\lambda^{(RA,RC)}$	0.1000	0.0926	−0.007	0.052
$\rho_\lambda^{(IC,RC)}$	0.1000	0.1057	0.006	0.056
$\sigma_{\lambda^{(RC)}}$	1.5000	1.5015	0.001	0.069
$\beta_0^{(Z)}$	−3.2792	−3.2504	0.029	0.200
$\beta_{IA}^{(Z)}$	0.5000	0.5214	0.021	0.080
$\beta_{RA}^{(Z)}$	0.5000	0.4946	−0.005	0.042
$\beta_{IC}^{(Z)}$	−0.5000	−0.5021	−0.002	0.076
$\beta_{RC}^{(Z)}$	−0.5000	−0.5009	−0.001	0.044

Web Appendix 6: Alternative Model Specifications and Performance

In this web appendix, we describe the modifications that were made to the GLS and SSW models to be able to incorporate time-varying covariates into them, and report detailed performance comparisons of the performance of the proposed method compared to these and other baseline methods.

We first consider augmenting the model of GLS with seasonal covariates. Recall that GLS models cumulative acquisitions and losses via non-linear least squares. We incorporate covariates into both the acquisition and the loss processes. We incorporate covariates into the acquisition process through proportional hazards.

The acquisition model of GLS has three parameters – α, β, and γ. α governs the total number of customers who will eventually be acquired in the future. Let the baseline without-covariates CDF characterizing the

Table 7 Parameter recovery analysis: proposed method, all parameters, additional parameter scenario 1

Parameter	True Param.	Est Param.	Bias	SD of Ests
$\lambda_0^{(IA)}$	0.1000	0.1005	0.000	0.023
$c^{(IA)}$	0.8000	0.8078	0.008	0.089
$\pi^{(IA)}$	0.9000	0.9004	0.000	0.011
$\lambda_0^{(RA)}$	0.1000	0.0990	−0.001	0.009
$c^{(RA)}$	0.8000	0.8011	0.001	0.021
$\pi^{(RA)}$	0.9000	0.8999	0.000	0.008
$\lambda_0^{(IC)}$	0.2000	0.1956	−0.004	0.021
$c^{(IC)}$	0.6000	0.6209	0.021	0.057
$\lambda_0^{(RC)}$	0.2000	0.2013	0.001	0.014
$c^{(RC)}$	0.6000	0.5986	−0.001	0.015
$\sigma_{\lambda^{(IA)}}$	2.0000	2.0238	0.024	0.326
$\rho_\lambda^{(IA,RA)}$	0.0000	0.0175	0.018	0.066
$\sigma_{\lambda^{(RA)}}$	1.0000	1.0137	0.014	0.066
$\rho_\lambda^{(IA,IC)}$	0.0000	−0.0052	−0.005	0.031
$\rho_\lambda^{(RA,IC)}$	0.0000	0.0001	0.000	0.061
$\sigma_{\lambda^{(IC)}}$	2.0000	2.0879	0.088	0.270
$\rho_\lambda^{(IA,RC)}$	0.0000	−0.0134	−0.013	0.079
$\rho_\lambda^{(RA,RC)}$	0.0000	0.0093	0.009	0.068
$\rho_\lambda^{(IC,RC)}$	0.0000	−0.0106	−0.011	0.117
$\sigma_{\lambda^{(RC)}}$	1.0000	1.0097	0.010	0.063
$\beta_0^{(Z)}$	−3.2992	−3.3795	−0.080	0.277
$\beta_{IA}^{(Z)}$	0.5000	0.5127	0.013	0.128
$\beta_{RA}^{(Z)}$	0.5000	0.4607	−0.039	0.104
$\beta_{IC}^{(Z)}$	−0.5000	−0.4988	0.001	0.080
$\beta_{RC}^{(Z)}$	−0.5000	−0.5044	−0.004	0.088

timing with which prospects are acquired be given by

$$F_0(t|\beta,\gamma) = \frac{1}{1+\exp(-\beta-\gamma\times t)}.$$

The corresponding CDF with time-varying covariates incorporated by proportional hazards is then

$$F(t|\beta,\gamma,\boldsymbol{\beta}_{cov}) = 1-\exp\left(-\int_0^t h(u|\beta,\gamma,\boldsymbol{\beta}_{cov})du\right), \qquad \text{where}$$

$$\int_0^t h(u|\beta,\gamma,\boldsymbol{\beta}_{cov})du = \sum_{i=1}^{t}\{\log[1-F_0(i-1)]-\log[1-F_0(i)]\}\exp(\boldsymbol{\beta}_{acq}^T\mathbf{x}(i)),$$

letting $\boldsymbol{\beta}_{cov}$ and $\mathbf{x}(i)$ denote the vector of coefficients associated with the time-varying covariates and the corresponding vector of covariate values in time period i, respectively.

Table 8 Parameter recovery analysis: proposed method, all parameters, additional parameter scenario 2

Parameter	True Param.	Est Param.	Bias	SD of Ests
$\lambda_0^{(IA)}$	0.0010	0.0011	0.000	0.000
$c^{(IA)}$	1.2000	1.1998	0.000	0.033
$\pi^{(IA)}$	0.9000	0.8812	−0.019	0.107
$\lambda_0^{(RA)}$	0.2000	0.2055	0.006	0.050
$c^{(RA)}$	0.6000	0.5954	−0.005	0.033
$\pi^{(RA)}$	0.9000	0.9091	0.009	0.022
$\lambda_0^{(IC)}$	0.1000	0.1137	0.014	0.051
$c^{(IC)}$	0.8000	0.7902	−0.010	0.078
$\lambda_0^{(RC)}$	0.2000	0.2140	0.014	0.052
$c^{(RC)}$	0.6000	0.6064	0.006	0.035
$\sigma_{\lambda^{(IA)}}$	2.0000	1.9693	−0.031	0.195
$\rho_\lambda^{(IA,RA)}$	0.0000	0.0043	0.004	0.145
$\sigma_{\lambda^{(RA)}}$	1.0000	1.0045	0.005	0.111
$\rho_\lambda^{(IA,IC)}$	0.2000	0.1828	−0.017	0.122
$\rho_\lambda^{(RA,IC)}$	0.2000	0.1942	−0.006	0.106
$\sigma_{\lambda^{(IC)}}$	2.0000	1.9767	−0.023	0.313
$\rho_\lambda^{(IA,RC)}$	0.2000	0.1786	−0.021	0.178
$\rho_\lambda^{(RA,RC)}$	0.2000	0.1306	−0.069	0.152
$\rho_\lambda^{(IC,RC)}$	0.0000	−0.0478	−0.048	0.144
$\sigma_{\lambda^{(RC)}}$	1.0000	1.0194	0.019	0.136
$\beta_0^{(Z)}$	−1.3539	−1.4036	−0.050	0.839
$\beta_{IA}^{(Z)}$	0.5000	0.5403	0.040	0.169
$\beta_{RA}^{(Z)}$	0.5000	0.4881	−0.012	0.288
$\beta_{IC}^{(Z)}$	−0.5000	−0.5634	−0.063	0.124
$\beta_{RC}^{(Z)}$	−0.5000	−0.5610	−0.061	0.273

The retention model of GLS is a single parameter, r, denoting the time-invariant per-period retention rate. We incorporate time-varying covariates into this process by modeling the logit of r as a function of time-varying covariates:

$$\text{logit}(r_t) = \boldsymbol{\beta}_{ret}^T \mathbf{x}(t)$$

Next we consider augmenting the model of SSW with seasonal covariates. Recall that unlike GLS, SSW models cumulative net (not gross) adds, which is the size of the customer base over time. We may add covariates to the cumulative net adds process via proportional hazards using exactly the same process that was used for the gross adds process for GLS. The churn rate in SSWs model is determined by taking a

Table 9 Parameter recovery analysis: proposed method, all parameters, additional parameter scenario 3

Parameter	True Param.	Est Param.	Bias	SD of Ests
$\lambda_0^{(IA)}$	0.0010	0.0011	0.000	0.000
$c^{(IA)}$	1.2000	1.1965	−0.003	0.030
$\pi^{(IA)}$	0.9000	0.8954	−0.005	0.085
$\lambda_0^{(RA)}$	0.2000	0.2257	0.026	0.066
$c^{(RA)}$	0.6000	0.6112	0.011	0.032
$\pi^{(RA)}$	0.9000	0.8921	−0.008	0.021
$\lambda_0^{(IC)}$	0.2000	0.2075	0.007	0.056
$c^{(IC)}$	0.6000	0.5928	−0.007	0.067
$\lambda_0^{(RC)}$	0.2000	0.2143	0.014	0.077
$c^{(RC)}$	0.6000	0.6018	0.002	0.031
$\sigma_{\lambda^{(IA)}}$	2.0000	1.9850	−0.015	0.198
$\rho_\lambda^{(IA,RA)}$	0.0000	−0.0626	−0.063	0.165
$\sigma_{\lambda^{(RA)}}$	1.0000	1.0089	0.009	0.128
$\rho_\lambda^{(IA,IC)}$	0.2000	0.2023	0.002	0.075
$\rho_\lambda^{(RA,IC)}$	0.2000	0.1684	−0.032	0.106
$\sigma_{\lambda^{(IC)}}$	2.0000	1.9460	−0.054	0.329
$\rho_\lambda^{(IA,RC)}$	0.2000	0.1763	−0.024	0.235
$\rho_\lambda^{(RA,RC)}$	0.2000	0.1770	−0.023	0.167
$\rho_\lambda^{(IC,RC)}$	0.0000	−0.0187	−0.019	0.174
$\sigma_{\lambda^{(RC)}}$	1.0000	1.0419	0.042	0.119
$\beta_0^{(Z)}$	−1.0129	−0.7712	0.242	0.972
$\beta_{IA}^{(Z)}$	0.5000	0.5798	0.080	0.196
$\beta_{RA}^{(Z)}$	0.5000	0.5806	0.081	0.389
$\beta_{IC}^{(Z)}$	−0.5000	−0.5698	−0.070	0.177
$\beta_{RC}^{(Z)}$	−0.5000	−0.5864	−0.086	0.232

trailing 12 month average empirical churn rate and projecting this rate forward. Given the lack of a formal model for the SSW retention process, we leave it as-is.

In Table 12, we summarize the predictive performance of the proposed method against the models from Gupta et al. (2004), Schulze et al. (2012), and McCarthy et al. (2017) (GLS, SSW, and MFH, respectively), as a function of the time horizon of the prediction in the rolling predictive analysis described in Section 6.1 of the main text.

Web Appendix 7: Publicly Disclosed Customer Data

Spotify's publicly disclosed customer data is provided in Table 13.

Table 10 Parameter recovery analysis: proposed method, all parameters, additional parameter scenario 4

Parameter	True Param.	Est Param.	Bias	SD of Ests
$\lambda_0^{(IA)}$	0.1000	0.1035	0.004	0.014
$c^{(IA)}$	0.8000	0.7937	−0.006	0.055
$\pi^{(IA)}$	0.9000	0.8998	0.000	0.012
$\lambda_0^{(RA)}$	0.1000	0.0980	−0.002	0.010
$c^{(RA)}$	0.8000	0.7981	−0.002	0.017
$\pi^{(RA)}$	0.9000	0.9012	0.001	0.008
$\lambda_0^{(IC)}$	0.2000	0.1994	−0.001	0.030
$c^{(IC)}$	0.6000	0.5969	−0.003	0.074
$\lambda_0^{(RC)}$	0.2000	0.1997	0.000	0.020
$c^{(RC)}$	0.6000	0.6018	0.002	0.018
$\sigma_{\lambda^{(IA)}}$	2.0000	1.9579	−0.042	0.194
$\rho_\lambda^{(IA,RA)}$	0.0000	0.0179	0.018	0.059
$\sigma_{\lambda^{(RA)}}$	1.0000	0.9973	−0.003	0.056
$\rho_\lambda^{(IA,IC)}$	0.2000	0.2160	0.016	0.030
$\rho_\lambda^{(RA,IC)}$	0.2000	0.2037	0.004	0.055
$\sigma_{\lambda^{(IC)}}$	2.0000	1.9929	−0.007	0.338
$\rho_\lambda^{(IA,RC)}$	0.2000	0.1840	−0.016	0.060
$\rho_\lambda^{(RA,RC)}$	0.2000	0.2130	0.013	0.053
$\rho_\lambda^{(IC,RC)}$	0.0000	0.0358	0.036	0.079
$\sigma_{\lambda^{(RC)}}$	1.0000	1.0372	0.037	0.059
$\beta_0^{(Z)}$	−2.9637	−2.8956	0.068	0.213
$\beta_{IA}^{(Z)}$	0.5000	0.5238	0.024	0.074
$\beta_{RA}^{(Z)}$	0.5000	0.5187	0.019	0.110
$\beta_{IC}^{(Z)}$	−0.5000	−0.5226	−0.023	0.104
$\beta_{RC}^{(Z)}$	−0.5000	−0.4773	0.023	0.066

Table 11 Parameter recovery analysis: proposed method, baseline scenario

	True selection mechanism		
	Linear	Square root	Quadratic
Mean absolute bias (% of true parameter)			
Heterogeneity means	1.0%	7.4%	8.6%
Heterogeneity variances	0.6%	4.2%	4.0%
Heterogeneity correlations	6.8%	32.5%	37.2%
Homogeneous parameters	0.2%	1.6%	1.5%
Selection parameters	1.4%		
Absolute coefficient of variation			
Heterogeneity means	11.0%	11.8%	13.4%
Heterogeneity variances	6.8%	7.4%	7.1%
Heterogeneity correlations	42.1%	44.8%	72.9%
Homogeneous parameters	3.2%	3.2%	3.2%
Selection parameters	10.9%		
Coverage rate (95% confidence interval)			
Heterogeneity means	95.0%	92.5%	62.5%
Heterogeneity variances	95.0%	87.5%	52.5%
Heterogeneity correlations	98.3%	81.7%	68.3%
Homogeneous parameters	95.0%	96.7%	73.3%
Selection parameters	98.0%		

Table 12 Spotify: Average MAPE by Forecasting Horizon for all Disclosures and Models

Disclosure	Horizon	GLS	SSW	MFH	AGG	MPL
ADD	1	15.8%	12.4%	6.2%	10.1%	9.7%
	2	23.1%	16.8%	9.6%	11.0%	9.6%
	3	25.7%	25.9%	13.3%	11.2%	12.3%
	4	24.3%	34.3%	16.3%	11.8%	11.2%
	5	31.6%	35.4%	22.5%	17.6%	19.6%
	6	27.8%	49.8%	35.8%	25.3%	25.3%
LOSS	1	54.0%	11.0%	6.9%	8.4%	6.3%
	2	21.9%	16.5%	5.1%	12.2%	7.7%
	3	9.3%	22.7%	8.3%	8.3%	10.9%
	4	12.0%	32.2%	10.1%	8.8%	9.4%
	5	42.3%	40.2%	15.5%	17.5%	11.3%
	6	51.3%	47.2%	14.0%	40.6%	20.3%
END	1	16.7%	2.7%	1.7%	3.7%	2.9%
	2	18.1%	4.5%	3.6%	6.0%	4.2%
	3	20.7%	6.6%	6.2%	5.7%	4.6%
	4	23.2%	8.4%	9.1%	6.1%	4.4%
	5	32.0%	10.6%	13.2%	7.3%	6.4%
	6	33.2%	14.0%	19.8%	15.4%	9.0%

Note: GLS, SSW, and MFH refer to the models from Gupta et al. (2004), Schulze et al. (2012), and McCarthy et al. (2017), respectively, where GLS and SSW have been extended to allow for time-varying covariates as described above. AGG refers to the proposed model, estimated off of aggregate disclosures. Proposed refers to the proposed model, estimated off of aggregate disclosures and credit card panel data using the MPL procedure from Section 3 of the main text.

Table 13 Spotify's Publicly Disclosed Customer Data (MM)

Date	*END*	*ADD*	*LOSS*	Date	*END*	*ADD*	*LOSS*
Jul 2010	0.5			Sep 2015	24		
Mar 2011	1			Dec 2015	28	9.9	5.9
Jun 2011	1.5			Mar 2016	30	8	6
Nov 2011	2.5			Jun 2016	36	12.9	6.9
Jan 2012	3			Sep 2016	40	11.3	7.3
Aug 2012	4			Dec 2016	48	15.9	7.9
Dec 2012	5			Mar 2017	52	12.3	8.3
Mar 2013	6			Jun 2017	59	16.8	9.8
May 2014	10			Sep 2017	62	13.3	10.3
Nov 2014	12.5			Dec 2017	71	19.2	10.2
Jan 2015	15			Mar 2018	75	14.3	10.3
Mar 2015	18			Jun 2018	83	19.9	11.9
Jun 2015	22			Sep 2018	87	16.2	12.2

Note: Most *END* and *LOSS* data were obtained from SEC filings. The remainder of the data was disclosed in an investor presentation (Statista: `https://www.statista.com/statistics/244995/number-of-paying-spotify-subscribers/`). *ADD* data was derived from the *END* and *LOSS* data. *ADD* and *LOSS* figures encompass the preceding three months (e.g., *ADD* for Sep 2018 denotes all customers acquired in July, August, and September 2018).

References

Eicker F (1967) Limit theorems for regressions with unequal and dependent errors. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 59–82 (Berkeley, CA: University of California Press).

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

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.

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

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. *Handbook of Econometrics* 4:2111–2245.

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

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

Train K (2009) *Discrete choice methods with simulation* (Cambridge university press).