THE UNIVERSITY OF CHICAGO

ESSAYS ON THE INDUSTRIAL ORGANIZATION OF FINANCIAL MARKETS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY
MARCO LOSETO

CHICAGO, ILLINOIS
AUGUST 2024

To Elena and Francesco, for giving me purpose and keeping me smiling

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES xi

ACKNOWLEDGMENTS xiv

ABSTRACT xv

1 NETWORK GAMES OF IMPERFECT COMPETITION: AN EMPIRICAL FRAMEWORK 1
- 1.1 Introduction 1
- 1.2 Contributions to the Literature 4
- 1.3 Demand 9
  - 1.3.1 Products 9
  - 1.3.2 Utility 9
  - 1.3.3 Individual Demand 10
  - 1.3.4 Aggregate Demand and Price Elasticities 10
- 1.4 Oligopolistic Competition 13
  - 1.4.1 Bertrand Competition 13
  - 1.4.2 Cournot Competition 17
  - 1.4.3 Bertrand vs Cournot 19
- 1.5 Multiproduct Firms 22
  - 1.5.1 Multiproduct Firm Problem 23
  - 1.5.2 Equilibrium price-cost Margins for Multiproduct Firms 24
  - 1.5.3 Market Definition and Mergers 25
- 1.6 Application: The US Automobile Industry 26
  - 1.6.1 Data 27
  - 1.6.2 The Empirical Nash-Bertrand Network Model 31
  - 1.6.3 Estimation Results 34
- 1.7 Conclusion 43
- 1.8 Appendix: Proofs 45
- 1.9 Equilibrium existence 52
- 1.10 Decomposition of Cournot Markups 53
- 1.11 Appendix: Additional Figures and Tables 57
- 1.12 Appendix: Simulation of the Bertrand Network game 59
- 1.13 Appendix: From linear-quadratic to discrete choice 62

2 PLAN MENUS, RETIREMENT PORTFOLIOS, AND INVESTORS' WELFARE 64
- 2.1 Introduction 64
- 2.2 Contributions to the Literature 72
- 2.3 Institutional Setting and Data Sources 76
  - 2.3.1 What is an employer-sponsored retirement plan? 76

2.3.2 Data . . . 78
2.4 Demand . . . 80
2.4.1 Sponsors' menu choice problem . . . 81
2.4.2 Funds' plan inclusion probabilities . . . 83
2.4.3 Investors' retirement portfolio problem . . . 87
2.4.4 Plan asset demand system . . . 89
2.5 Supply . . . 91
2.5.1 Nash equilibrium fees . . . 94
2.6 Demand Identification and Estimation . . . 98
2.6.1 Identification and estimation of sponsors' preferences . . . 98
2.6.2 Estimates of sponsors' preferences . . . 101
2.6.3 Identification and estimation of investors' preferences . . . 105
2.6.4 Estimates of investors' preferences . . . 108
2.7 Price-cost Margins and Fee Decomposition . . . 114
2.7.1 Recovering funds' price-cost margins . . . 114
2.7.2 Decomposition of equilibrium fees . . . 117
2.8 Counterfactuals . . . 118
2.8.1 Eliminating preference for affiliated funds . . . 121
2.8.2 Mandating the inclusion of low-cost options . . . 122
2.8.3 Capping funds' expenses . . . 124
2.9 Conclusions . . . 125
2.10 Appendix: Additional Figures and Tables . . . 127
2.11 Appendix: Model derivations . . . 142
2.12 Appendix: Turnover ratios & identification . . . 162
2.12.1 Funds' turnover ratios. . . . 162
2.12.2 Identification: turnover vs. performance . . . 163
2.12.3 How does turnover affect fees? . . . 167
2.13 Appendix: Nesting investors & sponsors preferences . . . 169
2.14 Appendix: 401(k) Lawsuits . . . 173

3 OLIGOPOLISTIC COMPETITION, FUND PROLIFERATION, AND ASSET PRICES . . . 177
3.1 Introduction . . . 177
3.2 Related Literature . . . 183
3.3 Model . . . 186
3.3.1 Household . . . 186
3.3.2 Mutual Funds . . . 188
3.3.3 Management Companies . . . 190
3.3.4 Financial Market . . . 192
3.3.5 Discussion of model assumptions . . . 193
3.4 Equilibrium . . . 195
3.4.1 Equilibrium definition . . . 195
3.4.2 Steady state definition and existence . . . 197

3.4.3 Numerical solution for the equilibrium path . . . 200
3.5 Data . . . 202
3.5.1 Data sources . . . 202
3.5.2 Data construction . . . 203
3.6 Model estimation . . . 205
3.6.1 Estimation procedure . . . 206
3.6.2 Results . . . 210
3.6.3 Counterfactuals and welfare analysis . . . 214
3.6.4 Asset pricing implications . . . 219
3.7 Conclusions . . . 225
3.8 Appendix: Derivations and Proofs . . . 227
3.9 Appendix: Figures . . . 231
3.10 Appendix: Tables . . . 233

REFERENCES . . . 236

LIST OF FIGURES

1.1 Binscatter of car prices against (unweighted) product Bonacich centrality after partialling out time fixed effects. . . . 30
1.2 Distribution of marginal costs. . . . 37
1.3 Distribution of price-cost margins. . . . 37
1.4 Decomposition of price-cost margins. All variables are measured in $1000 40
1.5 PCM/MPCM in Network Bertrand 41
1.6 Matrix of estimated own and cross price elasticities for the year 1990. The products included are the ones with an estimated own price elasticity above the median. 43
1.7 Matrix of estimated own and cross price elasticities for the year 1990. The products included are the ones with an estimated own price elasticity above the 25th percentile. 58
1.8 Simulated Network. Location is exogenous. Node size is proportional to markups. 60
1.9 Simulated Network. Price-cost margins (y-axis) against Bonacich centrality (x-axis). . . . 61

2.1 Plan quality for the year 2019. 'No S&P 500' is the share of plans without an S&P 500 tracker with a fee below 10bp. . . . 66
2.2 Distribution of asset-weighted and unweighted average plan expense ratio across plans for the year 2019. . . . 66
2.3 Binned scatter of plan (gross) performance and average plan expense corresponding to the specification in the second column of Table 2.5. Expenses and performance are yearly demeaned and measured in percentage points. . . . 67
2.4 Administrative structure of a defined-contribution employer-sponsored retirement plan. . . . 78
2.5 Cross-sectional estimates of plan investors median portfolio elasticity to funds' fees. Dashed vertical line corresponds to the DOL fee disclosure reform. . . . 112
2.6 Cross-sectional estimates of plan investors median portfolio elasticity to funds' fees. Dark blue squared-dashed line is the median sponsor elasticity. . . . 112
2.7 Cross-sectional estimates of the fraction of inactive investors (black). Share of plan investors in Vanguard plans holding a single TDF. . . . 114
2.8 Dollar change in surplus and average plan expense relative to the status quo for an investor holding a retirement account of $35,000 under different plan design policies. Magnitudes are in dollars-per-year. . . . 122
2.9 Distribution of number of options offered within investment category. . . . 127
2.10 Within in fund × year share of employers who meet minimum investment required for cheapest share class but offer a more expensive one. The black line is the average share of employers without cheapest share class. . . . 128
2.11 Distribution of average asset-weighted plan expense over time. Expenses are measured in percentage points 128
2.12 Distribution of plan expenses by plan size groups. Plan size is measured in number of participants. . . . 129

2.13 Median asset-weighted plan expense (dot-solid). Average expense ratio for a portfolio of Vanguard retail index funds (triangle-dashed). Expenses are measured in percentage points. 129
2.14 Within recordkeeper × six-digit NAICS dispersion in expenses. Other controls include plan assets, number of participants, and number of options. 130
2.15 Average overlap in recordkeepers' network of funds. A fund belongs to a recordkeeper's network if it is offered in a plan managed by that same recordkeeper. The red bars represent the average fraction of funds that belong to the network of any two of the 10 largest recordkeepers. The turquoise bars represent the asset-weighted overlap. 130
2.16 Distribution of number of options within category by plan size. 131
2.17 Distribution of number of options within category pre and post 2014. 131
2.18 Distribution of number of options within category by asset class. 131
2.19 Coefficient from regressing log(inclusion probability) on affiliation dummy. Inclusion probability is the share of 401(k) plans offering a given fund. Inclusion probabilities are computed at the (year × size group × industry × recordkeeper) level for each fund. Size groups are based on the number of plan participants. Industry is the 2-digit NAICS. 132
2.20 Share of retirement plans that offer at least one Target-Date-Fund (TDF). . . . 132
2.21 Average portfolio share across plan menus by asset class. Equity includes both US and International Equity funds. Balanced includes aggressive, moderate and conservative allocation funds that are not Target Date Funds (TDFs). 133
2.22 Secular decline in fees by fund type. The sample includes only funds available since 2010. The series for each type of fund has been shifted by the average fee as of 2010. 133
2.23 Secular decline in fees by fund type. The sample also includes funds introduced after 2010. The series for each type of fund has been shifted by the average fee as of 2010. 133
2.24 Cross-sectional decomposition of the average expense ratio into monopolist fee, hotelling markdown and plan inclusion markdown as defined in equation (2.20). Magnitudes are in basis points. 134
2.25 Expense ratio is residual expense ratio after absorbing funds' brand, year and category fixed effects. Instrument is residual turnover ratio. 165
2.26 Performance is yearly-demeaned alpha from 3 FF factors plus Momentum. Instrument is residual turnover. 165
2.27 Performance is yearly-demeaned (gross) return relative to BrightScope category return. Instrument is residual turnover. 165
2.28 Performance is yearly-demeaned (gross) return relative to Morning Star category return. Instrument is residual turnover. 165
2.29 Performance is next period yearly-demeaned (gross) return relative to BrightScope category return. Instrument is residual turnover. 166
2.30 Performance is next period yearly-demeaned (gross) return relative to Morning Star category return. Instrument is residual turnover. 166

2.31 Performance is next period yearly-demeaned alpha from 3 FF factors plus Momentum. Instrument is residual turnover. 166

3.1 Left Axis: AUM in trillions of $ for both passive and active equity industry. Right Axis: Share of AUM held in the passive industry. 179
3.2 Market share of the five biggest investment companies by investment strategy. Market shares are in terms of end-of-year assets under management (AUM). . . 180
3.3 Average number of funds per management company. Funds with different share classes count as a single fund. 181
3.4 Number of funds operated by each of the top five management companies over time. 211
3.5 Time-series of the average number of funds operated by the top five management companies as well as the number of funds operated by the outside management company in model vs data. In data, the time-series of the number of funds operated by the outside management company is computed as simple average of the number of funds operated by all non-top five management companies 213
3.6 Time-series of value-weighted fee from data and equilibrium fee estimated from the model. The value-weighted fee from data is estimated, for each year, by averaging the expense ratio reported by CRSP for each fund with weights proportional to lagged total net assets. 214
3.7 Estimated time-series of average revenues gained by the top five management companies and by the outside management company in the estimated model. . . 215
3.8 Percentage change in average number of funds, average fee, revenues and consumer surplus in each counterfactual compared to the model solution. For number of funds, we report the percentage change in average number of funds, where the average is computed across time. For the fee, we report the percentage change in the average fee, where the average is computed across time. For revenues, we report the percentage change in average revenues, where the average is computed across time and management companies. 216
3.9 Percentage change in average number of funds, average fee, revenues and consumer surplus in the counterfactual where Blackrock is replaced by two firms identical to Charles Schwab. For number of funds, we report the percentage change in average number of funds, where the average is computed across time. For the fee, we report the percentage change in the average fee, where the average is computed across time. For revenues, we report the percentage change in average revenues, where the average is computed across time and management companies. 218
3.10 Equilibrium comparative static with respect to initiation costs c. 221
3.11 Equilibrium comparative static with respect to dividend D. 223
3.12 Market share of the five biggest investment companies in the passive industry. Market shares are in terms of end-of-year assets under management (AUM). . . 231
3.13 Average number of passive funds per management company. Funds with different share classes count as a single fund. 232

3.14 Average asset-weighted fee across passive funds. Funds with different share classes count as a single fund. 232

LIST OF TABLES

1.1 Sales-weighted averages. 28
1.2 Demand estimates. Both specifications include interactions between characteristics and income, not reported here but available in Appendix 1.11. Standard errors are clustered at the car model level. 35
1.3 Sales-weighted averages of the decomposition of price-cost margins (PMC). 38
1.4 Demand estimates. Standard errors are clustered at the model level. 57
1.5 Parameters for simulation of Bertrand network game 59

2.1 Plan level summary statistics for the years 2010 to 2019. Each variable is first averaged within plan across years and then tabulated across plans. The variable 'N' is the number of plans and the variable 'N. of years' is the number of years a plan is observed in the sample. Sample is for sponsors with number of participants between 100 and 5000. 79
2.2 Two-step GMM estimates of plan sponsor preferences. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. 103
2.3 Estimates of plan investors preferences. All specifications include year, category and passive fixed effects. Expense ratios are in percentage points (pp.). R2 for IV columns is first stage. ME are the (median) marginal effects for portfolio allocations in pp. for a basis point increase in expenses or a pp. increase gross returns. 110
2.4 Price cost margins and marginal costs implied by the Nash-Bertrand first order conditions. Magnitudes are in basis points. 116
2.5 Fund performance is the difference between fund return and the average category return. Plan performance is the average performance (possibly asset weighted) of all funds in the plan. Returns are gross of fees. Returns and fees are in percentage points. 135
2.6 Dependent variable is funds' expense ratio. Independent variable is the number of funds within an investment category. Expense ratios are in percentage points. 135
2.7 BrightScope Beacon data coverage. Assets are in billions. 136
2.8 Fund level summary statistics for the years 2010 to 2019. Each variable is first averaged (or summed in the case of 'total assets') within fund-year across plans, then within plan across years and tabulated across funds. The variable 'N' is the number of funds, excluding cash accounts and company stocks. Portfolio share (sd.) is the within fund-year standard deviation of the fund portfolio share across plans, which is then averaged within fund across years. 137
2.9 Two-step GMM estimates of plan sponsor preferences. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. For the heterogeneous q specification, q varies at the year-recordkeeper-category level. 138

2.10 Two-step GMM estimates of plan sponsor preferences for plans with number of participants below the median (small) and above the median (large). Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. 138
2.11 Two-step GMM estimates of plan sponsor preferences for the pre 2014 and post 2014 subsamples. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. 139
2.12 Two-step GMM estimates of plan sponsor preferences accounting for inertia in menu choices. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. Sample is restricted to plans observed for at least two consecutive years. 139
2.13 Dependent variable is plan-level portfolio allocations. All specification include plan × year fixed effects. Beta are 3 Fama-French plus Momentum and 3 bond factors. 140
2.14 Estimates of plan investors preferences. All specifications include year, category and passive fixed effects. Expense ratios are in percentage points (pp.). R2 for IV columns is first stage. ME are the (median) marginal effects for portfolio allocations in pp. for a basis point increase in expenses or a pp. increase gross returns. Turnover and Hausman IV are standardized. 140
2.15 Decomposition of fees following equation (2.20). All magnitudes are in basis points. The figures shown are averages across time and funds. 140
2.16 Investor surplus and average plan expense under different counterfactual policies. Magnitudes are in basis points. Savings are relative to the status quo. Fee savings per-year assumes a retirement account balance of $35,000. Fee savings over 40 years assumes an annual income of $70,000, contribution rate of 10% and an annual return of 6%. Expense ratio cap is at 60 basis points. 141
2.17 Correlation table of instrument (turnover ratio) with fund performance measures. Turnover ratio and expense ratios are residuals after absorbing funds' brand, year and category fixed effects. Performance measures are yearly-demeaned. 164
2.18 Estimated investors and sponsors parameters assuming investors make their portfolio choice according to a discrete choice demand with linear random utility. . . 171

3.1 Calibrated inputs 206
3.2 Summary statistics and estimated inputs 210
3.3 Estimated parameters 211
3.4 Estimated multiplier 220
3.5 Summary statistics of the full sample. All variables are winsorized at 1% and 99% levels. Returns and alpha are monthly. The expense ratio is annual. 233
3.6 Summary statistics for the passive sample. All variables are winsorized at 1% and 99% levels. Returns, alpha and expense ratios are monthly. The expense ratio is annual. 233

3.7 Top 30 passive funds in the Large Cap sector. 234
3.8 Top 30 passive funds in the Mid Cap sector. 235

ACKNOWLEDGMENTS

I am deeply grateful to my advisors Lars Peter Hansen, Ali Hortaçsu and Scott Nelson for their unwavering guidance and support. Their invaluable insights and counsel have played a pivotal role in shaping each of the essays included in this dissertation and have significantly contributed to my growth as a researcher.

For the uncountable discussions and for introducing me to finance, I owe special thanks to Federico Mainardi.

For their friendship and support, I am grateful to Olivia Bordeu, Santiago Franco, Elena Istomina, Federico Mainardi, Pauline Mourot, Aleksei Oskolkov, Francesco Ruggieri and Marcos Sorá.

I gratefully acknowledged financial support from the Bradley fellowship, Stevanovich fellowship, Neubauer PhD fellowship and the Bank of Italy's Bonaldo Stringher scholarship. Part of the research appearing in this dissertation was also funded in part by the John and Serena Liew Fellowship Fund at the Fama-Miller Center for Research in Finance, University of Chicago Booth School of Business.

Last but not least, I am endlessly grateful to my family for their unconditional support throughout this journey, the journeys that have come to an end and the ones that are yet to come.

ABSTRACT

This dissertation consists of three essays. The first essay *"Network Games of Imperfect Competition: An Empirical Framework"* studies how product differentiation affects the ability of firms to price above costs in oligopolistic markets where products are differentiated over multiple attributes. It shows that, under suitable assumptions, firms' pricing power is fully summarized by a measure of network centrality function of product characteristics. The second essay *"Plan Menus, Retirement Portfolios and Investors' Welfare"* studies the US market of employer-sponsored retirment plans. It develops a structural model of plan menu design and fee competition between funds to rationalize the incentives behind the provision of high-cost investment options and to assess the effects of alternative plan design policies on investors welfare. The third essay *"Oligopolistic Competition, Fund Proliferation, and Asset Prices"* (joint work with Federico Mainardi) develops and estimates a dynamic oligopoly model to understand why different mutual fund families introduce new funds at different rates and at different points in time and how these competitive dynamics impact asset prices and investors welfare.

CHAPTER 1

NETWORK GAMES OF IMPERFECT COMPETITION: AN EMPIRICAL FRAMEWORK

1.1 Introduction

Product differentiation plays a crucial role in allowing firms to price above costs. When products are homogeneous, standard models of price competition predict that firms should price at costs even in markets with only two competitors.[1] The most common approach to model product differentiation is to define a product in terms of a bundle of characteristics. Although models of oligopolistic competition may differ in the way they are specified, they typically carry the same intuition about how product differentiation affects products' substitution patterns and producers' price-cost margins. Products that have similar characteristics should be more substitutable to each other, and firms selling more differentiated products should be able to charge higher margins. Although the intuition is clear, in models where the characteristic space is multidimensional, it is often hard to characterize analytically how product differentiation enables firms to price above costs.[2]

This paper considers a framework in which product differentiation enters substitution patterns and firms' price-cost margins in an way that is analytically tractable and easy to interpret. At the same time, it allows for products to be differentiated over multiple attributes and for consumers to have heterogeneous preferences over these attributes. In modelling consumer preferences, I make two assumptions. First, consumers choose how much to consume of each product. With this first assumption, I depart from the unit demand framework commonly used in empirical applications.[3] Second, product characteristics

1. Also assuming there are no capacity constraints.

2. This is not the case in more stylized models, such as the Hotelling and Salop models, where product are differentiated along one dimension.

3. However, if consumers were constrained to purchase at most one unit of a single product, the model

enter consumers' preferences in a linear-quadratic fashion. The linear component of preferences mimics the linearity of consumers' indirect utility, often assumed under discrete choice demand. The quadratic term instead enters preferences with a negative sign, implying that consumers have a taste for variety. This preference specification induces an hedonic demand system similar to the generalized hedonic-linear (GHL) demand system recently developed in Pellegrino (2023).[4]

Under the previous assumptions, I show that the cross-price elasticity between any two products is proportional to a weighted inner product of their corresponding vector of attributes. This weighting has two properties. First, it is such that, regardless of the units with which characteristics are measured, the resulting inner product always lies in between -1 and 1 and thus can be interpreted as a correlation between the characteristics of the two products. Two products are substitutes if their characteristics are positively correlated and complements otherwise. Second, the weight on a given characteristic is inversely related to how much that characteristic varies across all products in the market. Thus, characteristics that are more homogeneous across all products available matter more in determining substitution patterns between products.

Next, I turn to the supply side to study how product attributes affect firms' pricing power. Although I will primarily focus on Bertrand competition because it is the workhorse model used in empirical applications, I also consider quantity competition a la Cournot. In both cases, I leverage the linear-quadratic structure of consumer preferences to frame the oligopolistic game as a network game in the spirit of Ballester, Calvó-Armenagol and Zenou (2006). Products are the network nodes, and the weighted inner product between their vector

would boil down to a standard discrete choice demand framework. The linear-quadratic model nests the discrete choice framework. The only difference would be that the inner-product of the vector of attributes of, say, product j, would enter the indirect utility u_{ij} as an additional product characteristic. For more details see Appendix 1.13.

4. In studying the demand and supply of differentiated products under perfect competition, Epple (1987) also leverages these type of preferences to study the identification of hedonic demand systems of the type considered in the seminal work by Rosen (1974)

of attributes determines the strength of the links between nodes. The implied competitive network is weighted, undirected and such that the more substitutable two products are, the stronger their link. Within this network framing, I show that firms' equilibrium price-cost margins can be decomposed additively into two components: a monopolistic component and a product differentiation component. The monopolistic component captures the price-cost margins a monopolist would charge. The product differentiation component instead summarizes how differentiated a product is relative to its competitors and is proportional to the Bonacich network centrality of that product. This centrality measure enters firms' price-cost margins with a negative sign because a more central product or, equivalently, a less differentiated product faces more competition and, in turn, charges lower markups.

In the second part of the paper, I estimate the Bertrand Network model using price-quantity data on the US automobile industry. I show that the demand parameters can be identified and estimated with a simple linear IV strategy, assuming that any unobserved product characteristic enters consumers' utility only through the linear component of preferences. With the estimated demand parameters, margins and costs can be recovered from the supply equation. Although the model is not based on discrete choice demand, it delivers reasonable price elasticities and price-cost margins. When I decompose the estimated margins, I find that the network structure implied by the observed product characteristics is competitive. The Bonacich product centrality, on average, accounts for more than 90% of the monopolistic margins. By differentiating their products, car producers are able to capture from 2% to 7% of the potential monopolistic margins.

Lastly, to quantify how much product differentiation matters in determining price-cost margins, I compare the estimated margins with the ones firms would have charged if their products were to be homogeneous. The estimated margins can be as high as three times the homogeneous margins, suggesting that product differentiation plays an important role in allowing firms to price above costs.

The rest of the paper proceeds as follows. Section 1.2 describes how this paper fits the literature, Section 1.3 develops the demand side of the model, Section 1.4 focuses on the supply side and looks at both Bertrand and Cournot competition, Section 1.5 extends the supply to the case of multiproduct firms, Section 1.6 estimates the model with market level data on the US automobile industry and Section 1.7 concludes.

1.2 Contributions to the Literature

This paper contributes to the empirical industrial organization literature that estimates oligopolistic models of product differentiation using market-level data. Most of this literature models demand as a discrete choice problem and supply as a game of imperfect competition with differentiated products where, in most cases, firms are assumed to choose prices simultaneously (i.e., a la Bertrand).[5] Bresnahan (1987) was among the first to estimate a discrete choice model of oligopolistic competition with products that are vertically differentiated along one dimension (i.e., a la Hotelling). A few years later, motivated by the theoretical advancements in the modelling of product differentiation, Feenstra and Levinsohn (1995) extended the Bresnahan (1987) model to accommodate for product differentiation along multiple dimensions and showed how to recover price-cost margins in this more general context. Importantly, in their model, substitution patterns and markups are determined by the distance in the characteristic space to neighbouring products.

In parallel, Berry (1994) and Berry, Levinsohn and Pakes (1995) developed methodologies to estimate oligopolistic models in which products are horizontally differentiated. These methodologies can accommodate the presence of unobserved product characteristics, which is one of the reasons that made them the leading approaches to estimating oligopolistic models of product differentiation.[6] An important caveat with these models is that to obtain

5. For an exception with continuous demand see Dubois, Griffith and Nevo (2014).

6. Perhaps even more importantly, these methodologies enable researchers and practitioners to test firms'

reasonable substitution patterns, in the sense that products with similar characteristics are more substitute to each other, one needs to introduce random coefficients on product characteristics and estimate the model via a non-convex optimization. In addition, while in most cases, the estimates obtained from these models match the common intuition that similar products are more substitutable, it is not immediate to confirm such intuition analytically from the model equations.

In the framework I consider, substitution patterns and price-cost margins are related to the distance between product characteristics as in Feenstra and Levinsohn (1995) but with the advantage that the characterization is analytical instead of being defined implicitly. More precisely, by framing oligopolistic competition as a network game, I show that the equilibrium price-cost margins can be expressed as a function of a product's Bonacich centrality, which captures how close that product is to its competitors and represents a summary statistic for the ability of firms to price above marginal costs. I also show that the cross-price elasticity between any two goods is determined by a weighted inner product of their vector of characteristics which, differently from the standard discrete-choice models, can accommodate the presence of complementary goods.[7] At the same time, the model allows for unobserved product characteristics, delivers reasonable substitution patterns and can be estimated with a simple linear IV strategy.

Product proximity in terms of characteristics also matters for more practical aspects of model estimation. In the context of the standard discrete choice framework developed by Berry (1994) and Berry, Levinsohn and Pakes (1995), a recent contribution by Gandhi and Houde (2023) shows that, under the common assumption that product characteristics are

conduct on the the supply side (Nevo (2001)) and quantify the welfare effects of mergers.

7. A notable exception is Gentzkow (2007) who develops a discrete choice model with complementary goods. Departing from discrete choice, Thomassen, Smith, Seiler and Schiraldi (2017) develop and estimate a demand model of complementary product categories, where consumers have quadratic preferences. Similarly, Lee and Allenby (2009) study the role of complementarity across product categories. In their demand model, consumers' utility is a composite function with linear sub-utility within each product category, and a quadratic specification across categories.

exogenous, relevant price instruments should reflect the degree of differentiation of a product relative to others available in the market. In particular, they show that the residual function of the model, which depends on endogenous prices, is a function of the distances between observed product characteristics. In the model considered here, the intuition of this result emerges clearly from the analytical solution of the equilibrium prices (i.e., Proposition 3) which are a function of the proximity of products in the characteristic space as measured by their Bonacich network centrality.

Also recently, Magnolfi, McClure and Sorensen (2022) showed how to incorporate online survey data on the product space into demand estimation. The data capture product distances in the form of "product A is closer to B than it is to C" and can be used to construct a low-dimensional representation of the latent product space (i.e., an embedding). The authors show that incorporating these embeddings into conventional random coefficient logit models delivers elasticity estimates that are similar to those from a model that uses observable characteristics, suggesting that what matters are not attributes per se but rather how a product's attributes differ relative to its competitors. Similarly, in the network model developed here, substitution patterns depend only on how similar product characteristics are. Any set of characteristics that preserves the same network structure would generate the same substitution patterns.

This paper also contributes to a recent literature that applies results from network theory to study oligopolistic competition and market power. This literature builds on the seminal contribution by Ballester, Calvó-Armenagol and Zenou (2006) to frame games of imperfect competition as network games with product differentiation. In a recent contribution, Ushchev and Zenou (2018) develop a model of price competition in which product varieties are differentiated over a network and show that a unique Bertrand-Nash equilibrium exists and is proportional to a sign-alternating version of the Bonacich centrality. Galeotti, Golub, Goyal, Talamàs and Tamuz (2022) also frame oligopolistic price competition as a network

game and exploit properties of the singular value decomposition (SVD) of the Slustky matrix to characterize optimal tax-subsidy designs in terms of properties of the underlying network structure. Pellegrino (2023) instead considers a general equilibrium Cournot oligopoly in which products are differentiated over multiple attributes and the network structure is pinned down by the characteristics of the products.[8] He decomposes Cournot markups into a (quality-adjusted) productivity component and a product centrality component and calibrates the model using data on firms' financials from Compustat while measuring the adjacency matrix of product characteristics using the data on product cosine similarities constructed in Hoberg and Phillips (2016).

This paper contributes to this network literature along several dimensions. First, as in Pellegrino (2023), but differently from Ushchev and Zenou (2018) and Galeotti, Golub, Goyal, Talamàs and Tamuz (2022), I assume that products are differentiated over multiple attributes, which imply that the characteristics of the products determine the network structure. Second, differently from Pellegrino (2023), I focus on Bertrand competition, the standard conduct assumption made in most empirical applications. Importantly, I show that in both Cournot and Bertrand games, equilibrium price-cost margins can be decomposed additively into a monopolistic component that captures the margin a monopolist would charge and into a product differentiation component which captures how differentiated a given product is relative to its competitors.[9] I show that, in both Cournot and Bertrand, this centrality component coincides with the standard Bonacich network centrality as defined in Bonacich

8. See also Ederer and Pellegrino (2022) which extends the Cournot network model of Pellegrino (2023) to account for firms' institutional ownership structure.

9. The decomposition I obtain is remindful of the one obtained by Chen, Zenou and Zhou (2018) and Chen, Zenou and Zhou (2022) in a model of price competition with perfect price discrimination where firms sell products that generate network externalities between connected consumers. In that setting, the price charged to a given consumer decomposes additively into the monopoly price and the consumer network centrality. The latter enters negatively suggesting that firms' offer price discounts to more connected consumers. In the model considered here, firms cannot price discriminate and there are no network externalities. The network structure arises because products are differentiated over multiple attributes and more central products must charge lower prices because they are not differentiated enough.

(1987) and Jackson (2008) and, in both cases, the relevant adjacency matrix is a simple (possibly weighted) inner-product between the matrix of product characteristics X.[10] Moreover, similarly to Galeotti, Golub, Goyal, Talamàs and Tamuz (2022), I exploit the SVD of the Slusky matrix to characterize and interpret own and cross-price elasticities (Proposition 2).

Overall, this paper provides a unified framework to model imperfect competition in quantities or prices as a network game in which products are differentiated over multiple attributes, and shows how to estimate the model with market-level data (e.g., prices, quantities and characteristics) on a given industry using a simple linear IV strategy.

Finally, this paper contributes to the theoretical industrial organization literature that compares equilibrium outcomes across price and quantity competition. In general, this literature finds that under strategic complementarity in prices, Cournot equilibrium prices are higher than Bertrand prices, thus confirming the intuition that Bertrand competition is more intense. The seminal contribution by Singh and Vives (1984) develops the argument for the case of a differentiated duopoly which was then extended to an N firms oligopoly in Vives (1985).[11] More recently, Magnolfi, Quint, Sullivan and Waldfogel (2022) pointed out that, under the assumption that prices are strategic complements, imposing Cournot competition would always lead the researcher to estimate higher markups. In Proposition 5 I show that these results also hold when comparing the Cournot network game with the Bertrand network game. In particular, I show that, under strategic complementarity in prices, the centrality of each product is higher under Bertrand competition.

10. The decomposition I propose here is for the dollar price-cost margins $p - c$. For the Cournot case, Pellegrino (2023) instead develops a decomposition for markups defined as $\mu = p/c$. In Appendix 1.10, I show that the two decompositions are related. In particular, I show that the measure of centrality defined in Pellegrino (2023) is an affine transformation of the Bonacich product centrality that enters Cournot price-cost margins.

11. Similar results can be found in Cheng (1985), Okuguchi (1987) and Amir and Jin (2001).

1.3 Demand

This section sets up the demand side of the model. I start by describing the set of products available, consumers' preferences and by deriving individual demand functions. Then, I turn to aggregate demand and describe how product attributes affect own and cross price elasticities.

1.3.1 Products

In the market I consider, there are $j \in \{1, ..., J\}$ products available. Each product j is characterized by a set of K attributes whose values are collected in the K dimensional real-valued vector $x_j = (x_{jk})_{k=1}^K$ where x_{jk} is measured in units of quantity consumed. Characteristic x_{jk} tells you how much of attribute k you would get if you consume one unit of product j.

1.3.2 Utility

Consumers are indexed by $i \in I$, take product prices $p = (p_j)_{j=1}^J$ as given, and choose how much to consume of each product available. I denote by $q_i = (q_{ij})_{j=1}^J$ the consumption vector of consumer i.

I define consumer i preferences as follows:[12]

$$\begin{aligned} u_i(q_i, X) &= q_{i0} + \left(q_i' \alpha_i^q - \frac{\beta_i}{2} q_i' q_i \right) + \eta \left(q_i' X \alpha_i^x - \frac{\beta_i}{2} q_i' X X' q_i \right) \\ &= q_{i0} + q_i' \alpha_i - \frac{\beta_i}{2} q_i' \left(I_J + \eta X X' \right) q_i \end{aligned} \tag{1.1}$$

where q_{i0} is an outside good, X the $J \times K$ matrix of products attributes, $\eta > 0$ governs the extent with which product differentiation in terms of attributes matters for consumers,

12. The same type of preferences are considered in Pellegrino (2023).

$\alpha_i^q > 0$ and $\alpha_i^x > 0$ are respectively a J-vector and K-vector of utility parameters. Both vectors of parameters affect the marginal utility that comes from the linear term in (1.1), $\alpha_i \equiv \eta X \alpha_i^x + \alpha_i^q$. Finally, β_i captures i's love for varieties.

1.3.3 Individual Demand

Consumer i takes prices p as given and maximizes (1.1) subject to

$$q_{i0} + q_i' p \leq y_i \tag{1.2}$$

where y_i is consumer i income. After substituting for the budget constraint in (1.2), consumer i's demand function is given by

$$q_i(p) = \frac{1}{\beta_i} \left(I_J + \eta X X'\right)^{-1} (\alpha_i - p) \tag{1.3}$$

which is always well defined because $\left(I_J + \eta X X'\right)$ is positive definite and therefore non-singular.

1.3.4 Aggregate Demand and Price Elasticities

I assume there is a mass M of consumers that are heterogeneous in terms of (α_i, β_i). Further, I assume that all moments involving the random variables α_i and β_i are well-defined. The next proposition characterizes the aggregate demand function.

Proposition 1. *Under the above assumptions, the aggregate demand of product j as a function of own and competitors prices is given by*

$$q_j(p_j, p_{-j}) = a_j(\alpha, \beta, (\theta_{jl})_{l=1}^J) - \frac{1}{\beta}(1 - \theta_{jj})p_j + \frac{1}{\beta}\sum_{l \neq j} \theta_{jl} p_l \tag{1.4}$$

where $\beta \equiv M\left(\int \frac{1}{\beta_i} di\right)^{-1}$, $\theta_{jl} \equiv x_j' \Omega^{-1} x_l$ *where* $\Omega \equiv \left(\frac{1}{\eta} I_K + X'X\right)$ *is a positive definite weighting matrix independent of* (j,l) *and* a_j *is a* j*-specific demand intercept that depends on* $\alpha \equiv \int \frac{\alpha_i/\beta_i di}{\int 1/\beta_i di}$. *Moreover, for any* (j,l), $\theta_{jl} \in (-1,1)$ *if* $j \neq l$ *whereas* $\theta_{jj} \in (0,1)$ *if* $j = l$.

Equation (2.10) defines the aggregate demand for a given product j. Because consumer preferences are quadratic, aggregate demand is linear in prices. What is more interesting is how own and cross price elasticities relate to products characteristics. This relationship is enclosed in the elements $(\theta_{jl})_{j,l}$ which determine the substitution patterns across products. In fact, the (j,l) element of the Slutsky matrix is given by

$$\frac{\partial q_j}{\partial p_l} = \begin{cases} -\frac{1}{\beta}(1 - \theta_{jj}) & \text{if } j = l \\ \frac{1}{\beta}\theta_{jl} & \text{if } j \neq l \end{cases} \tag{1.5}$$

so that products j and l are substitutes or complements whenever θ_{jl} is respectively positive or negative. But Proposition 1 tells us more; the element θ_{jl} is a weighted inner-product between j and l vectors of attributes. In other words, the closer j and l are in this inner-product space the more substitutable they are. The next result characterizes the weighting of the inner-product between x_j and x_l in terms of the principal components of the matrix of characteristics X.

Proposition 2. *Let* U *be the* $K \times K$ *matrix of principal components directions of* X *and for any* j *let* $\tilde{x}_j = U'x_j$ *be the projection of* x_j *onto these principal components. Then for any* (j,l)

$$\theta_{jl} = \sum_{k=1}^{K} \left(\frac{1}{1 + \lambda_k^{x'x}}\right) \tilde{x}_{jk}\tilde{x}_{lk} \tag{1.6}$$

where $\lambda_k^{x'x}$ *is the* k*-th eigenvalue of* $X'X$.

Equation (1.6) has two main insights. First, the substitution between any two products j and l can be expressed as a weighted inner-product between the vectors of projected attributes $\tilde{x}_j$ and $\tilde{x}_l$. Because U is an orthogonal matrix, this projection is innocuous in the sense that $\tilde{X}\tilde{X}' = XUU'X' = XX'$ and we can replace XX' with $\tilde{X}\tilde{X}'$ without affecting consumer preferences defined in (1.1), individual demand defined in (1.3) and aggregate demand defined in (2.10).

Second, the weighting of these (projected) product characteristics depends on how much variety in terms of each characteristic is available in the whole market. From expression (1.6) we can see that characteristics with smaller $\lambda_k^{x'x}$ are weighted more. But what does a small $\lambda_k^{x'x}$ mean in practice? It means that higher weight is assigned to (projected) characteristics that do not vary too much across products as measured by $\tilde{X}'\tilde{X}$. To see this more formally, let u_k be k-th principal component of X, $\tilde{x}_k$ the k-th column of $\tilde{X}$ and note that,

$$\tilde{x}_k'\tilde{x}_k = u_k'X'Xu_k = u_k'U\Lambda^{x'x}U'u_k = \lambda_k^{x'x}. \tag{1.7}$$

where the first equality comes from the definition of $\tilde{x}_j$ and $\tilde{x}_k$, the second from the eigen-decomposition of $X'X$ and the last one from the fact that U is an orthogonal matrix. Overall, equation (1.7) tells us that characteristics that vary more across all products available in the market will have a higher $\lambda_k^{x'x}$ and thus will matter less when computing substitution patterns between any two products.

To sum up, equation (1.6) highlights that the elasticity of substitution between two products is affected by not only how similar are their vector of characteristics, but also by how similar are characteristics across the whole market. The substitutability between any two products will be higher if their characteristics are similar but even more so if the characteristics in which they are similar are the ones that are more homogeneous across all products available.

1.4 Oligopolistic Competition

In this section, I turn to the analysis of the supply side. I will mainly focus on Bertrand competition because it is the workhorse model used in empirical applications. After solving the Bertrand game, I will turn to Cournot competition. In both cases, I show that the effect of product differentiation on equilibrium markups is summarized by a measure of how central a product is in the competitive network. I conclude the section by comparing the two cases and by showing that, under strategic complementarity, Cournot competition always leads to higher price-cost margins.

To start with, I assume that J single-product firms produce the J products with constant marginal costs. Then, section 1.5 deals with the case in which firms are multiproductuct. Moreover, throughout the following analysis, I will assume that an interior Nash equilibrium exists.[13]

1.4.1 Bertrand Competition

Firm j takes the vector of competitor prices p_{-j} as given and solves

$$\max_{p_j} \quad (p_j - c_j) q_j(p_j, p_{-j}) \tag{1.8}$$

$$\text{s.t.} \quad q_j(p_j, p_{-j}) = a_j - \frac{1}{\beta}(1 - \theta_{jj}) p_j + \frac{1}{\beta} \sum_{l \neq j} \theta_{jl} p_l \tag{1.9}$$

which is equivalent to

$$\max_{p_j} \left(a_j + \frac{c_j}{\beta}(1 - \theta_{jj}) \right) p_j - \frac{1}{\beta} \left(1 - \theta_{jj}\right) p_j^2 + \frac{1}{\beta} \sum_{l \neq j} \theta_{jl} p_j p_l. \tag{1.10}$$

13. In Appendix 1.9, I provide conditions for the existence and uniqueness of an Nash equilibrium where $p_j \in [c_j, \alpha_j]$ for all j.

The payoff function in equation (1.10) is analogous to the linear-quadratic utility functions considered in Ballester, Calvó-Armenagol and Zenou (2006) and, as such, defines a linear-quadratic network game in which each product is a node and the $J \times J$ matrix

$$A(\Theta) \equiv \Theta - diag(\Theta) \tag{1.11}$$

is the weighted and undirected adjacency matrix of the network.

Network game interpretation. The adjacency matrix defined in (1.11) shows that network connections and products' substitution patterns are isomorphic to each other. From the previous section, we know that an off-diagonal element θ_{jl} of the matrix Θ captures the degree of substitution between product j and product l as measured by a weighted inner-product of their vector of characteristics x_j and x_l respectively. From equation (1.11), we know that we can interpret the inner-product θ_{jl} as a weighted link between product j and product l and therefore, we can think of the product differentiation space as being a network whose nodes are the products and whose links tell us how close, or equivalently how substitutable, any two products are.

The natural question at this point is, why embedding product differentiation into a competitive network is relevant? Framing the product differentiation space as a network is important because it enables us to learn how product differentiation affects equilibrium outcomes by studying the topological properties of the network. In the next proposition, I show that the equilibrium Bertrand price-cost margins depend negatively on a product's Bonacich centrality, which, following Jackson (2008), I define as

Definition 1. *Let (A, J) be a network with J nodes and adjacency matrix A. The J-vector of (weighted) Bonacich centralities $\mathbf{b}(A, \delta, u)$ is given by*

$$\mathbf{b}(A, \delta, u) \equiv (I_J - \delta A)^{-1} \delta A u = \sum_{k=1}^{\infty} \delta^k A^k u, \tag{1.12}$$

where $\delta > 0$ is a scalar and $u > 0$ is J-vector.

The j-th element of $\mathbf{b}(A, \delta, u)$ summarizes how central node j is in the network. This measure of centrality is widely used in social networks because it captures a node's importance in terms of how close/connected this node is to others and how close/connected the nodes it is connected to. According to the definition of Bonacich centrality, a node's importance is a weighted sum of the walks that emanate from it. Moreover, if $\delta \in (0, 1)$, walks of shorter length are weighted more.[14]

Proposition 3. *Assume that $\theta_{jj} = \theta$ for all $j \in \{1, ..., J\}$. If an interior equilibrium $p^* = (p_j^*)_{j=1}^J$ of the Bertrand pricing game exists, then it is unique and is such that*

$$p^* - c = \frac{\alpha - c}{2} - \mathbf{b}\left(A(\Theta), \frac{1}{2(1-\theta)}, \frac{\alpha - c}{2}\right), \tag{1.13}$$

provided $\theta < 1 - \frac{1}{2}\max_j |\lambda_j(A)|$ where $\lambda_j(A)$ is the j-th eigenvalue of the adjacency matrix $A(\Theta)$.

The key insight of Proposition 3 is that the more central a product is in the competitive network, the lower its equilibrium price-cost margins.[15] What does this mean in practice? From Definition 1, we can see that the higher any of the entries of the j-th row of A, the more central node j is. In our settings, product j is more central the higher its substitutability with any other product (i.e., the higher the elements $(\theta_{jl})_{l \neq j}$ of the j-th row of Θ). Overall, the expression for the Bertrand price-cost margins in (1.13) tells us two things. First, a less central or, equivalently, more differentiated product will be able to charge higher markups. Second, a product's Bonacich centrality is a sufficient statistic to measure how product differentiation allows firms to price above marginal costs.

14. This interpretation is motivated by the fact that when A is a binary $\{0, 1\}$ it k-th power A^k counts how many walks of length k are between any two nodes.

15. In Appendix 1.12 I perform a simple simulation exercise to visualize and summarize the properties of the Bertrand network model.

Comparison with Ballester, Calvó-Armenagol and Zenou (2006). In their seminal paper, Ballester et al. show that in a general network game with quadratic payoffs, the Nash equilibrium action of any player is increasing in their Bonancich centrality. For the Bertrand competition game I study, the opposite holds; Nash equilibrium prices decrease with a player's centrality. This mismatch in the results is a consequence of the fact that in Ballester et al., the coefficient on the linear component of players' utility does not depend on the network structure, whereas, in the Bertrand case, the linear term of the quadratic profit in (1.10) depends on Θ.[16]

What is the economic interpretation of this result? The linear term in the Bertrand game corresponds to the marginal benefit of the very first unit for the average consumer and (assuming $c_j = 0$) is given by

$$a_j = (1 - \theta_{jj})\alpha_j - \sum_{l \neq j} \theta_{jl}\alpha_l. \tag{1.14}$$

From expression (1.14), it is immediate to see that a more central product faces a lower residual demand. This *residual-demand* effect has to be contrasted with the *peer-effect* benefit of being more connected, which affects firms' payoffs through the interaction terms in

$$\sum_{l \neq j} \theta_{jl} p_l p_j. \tag{1.15}$$

From expression (1.15), we can see that holding everything else constant, being more central or having more substitutes increases profits. In the class of network games studied by Ballester et al., only this second effect is present, and it is the force that pushes players to

16. As I discuss in subsection 1.4.2, this is not the case with quantity competition a la Cournot. Consequently, the relationship between the Cournot-Nash equilibrium and the Bonacich network centrality follows immediately from the main result in Ballester, Calvó-Armenagol and Zenou (2006).

increase their equilibrium action proportionally to their network centrality. In the Bertrand network game, not only the *residual-demand* effect in (1.14) is present but, as I show in Proposition 3, it dominates the *peer-effect*. As a result, the equilibrium actions (i.e., firms' prices) decrease with a product's centrality.

Some remarks. A couple of remarks are in order before turning to the Cournot case. First, the assumption that θ_{jj} are homogeneous across j is made only for expositional purposes, and the proof presented in Appendix 1.8 is provided for the general heterogeneous case. Similarly, Section 1.5 generalizes the result to the case in which firms are multiproduct.

Second, the condition on the largest eigenvalue of $A(\Theta)$ ensures that the Bonacich centrality can be rewritten as a convergent infinite sum. If the condition is not satisfied, expression (1.13) would still be well-defined but not expressable as an infinite series.

Third, Proposition 3 offers a two-terms decomposition of price-cost margins, a monopolistic component and a product differentiation component summarized in terms of the network centrality. Interestingly, this latter component matters only to the extent that firms are competing with each other. As I show in Section 1.5, when there is a multiproduct monopolist, the product differentiation component converges to zero, and monopolistic price-cost margins are charged to each of the J product varieties.

1.4.2 Cournot Competition

In this section, I focus on quantity competition a la Cournot. To define a firm's objective, I need the inverse aggregate demand instead of the aggregate demand derived in Proposition 1. Compared to Bertrand, deriving the aggregate demand for the Cournot case is almost immediate. Starting from (1.3), and defining α and β as in Proposition 1 the aggregate

inverse demand for product j is

$$p_j(q_j, q_{-j}) = \alpha - \beta(1 + \eta(x_j' x_j))q_j - \eta\beta \sum_{l \neq j} (x_j' x_l) q_l. \tag{1.16}$$

Similar to the Bertrand case, substitution patterns in Cournot are driven by an inner product between product attributes. In this case, the inner product is unweighted, and the substitutability between products j and l is given by $x_j' x_l$. One drawback of this is that, without a normalization on the scale of characteristics, the implied product adjacency matrix can have weights larger than one in absolute value, which makes it less interpretable. In the Bertrand case, as shown in Proposition (3), no normalization of the product characteristics is required.

Firm j takes competitors' quantities as given and solves

$$\max_{q_j} \quad (\alpha_j - c_j)q_j - \beta(1 + \eta(x_j' x_j))q_j^2 - \eta\beta \sum_{l \neq j} (x_j' x_l) q_l q_j \tag{1.17}$$

The Cournot objective in (1.17) also defines a network game with quadratic payoffs. In this case, the relevant adjacency matrix will be constructed from the matrix $\Theta^- \equiv -XX'$. Why this is important will be clear in the following Proposition, where I derive the Cournot price-cost margins as a function of the vector of Bonacich centralities.

Proposition 4. *Let $\Theta^- \equiv -XX'$ and assume that $\theta^-_{jj} = \theta^-$ for all $j \in \{1, ..., J\}$. If an interior equilibrium $q^* = (q_j^*)_{j=1}^J$ of the Cournot game exists, then it is unique and is such that*

$$p^* - c = \frac{\alpha - c}{2} + \mathbf{b}\left(A(\Theta^-), \frac{\eta}{2(1 - \eta\theta^-)}, \frac{\alpha - c}{2}\right) \tag{1.18}$$

provided $\theta^- < \frac{1}{\eta} - \frac{1}{2}\max_j |\lambda_j(A(\Theta^-))|$ where $\lambda_j(A)$ is the j-th eigenvalue of the adjacency matrix $A(\Theta^-)$.

The expression for Cournot price-cost margins resembles very closely the one for Bertrand described in (1.13). The most apparent difference is that now price-cost margins seem to increase in the product's centrality, but this is not the case. To see this, note that the relevant adjacency matrix here is $\Theta = -XX'$, which, because of the minus sign, assigns higher centrality whenever a product becomes less substitutable with any other product or equivalently more differentiated. As expected, Cournot and Bertrand's price-cost margins are higher for more differentiated products.

In the context of Cournot competition, Pellegrino (2023) also develops a decomposition of firms' markups, defined as p_j/c_j, into a productivity component α_j/c_j and a product centrality component, denoted by $1-\chi_j$ in the paper. Although different, the two decompositions are related to each other. More precisely, in Appendix 1.10, I show that the product centrality defined in Pellegrino (2023) is an affine transformation of the Bonacich product centrality that enters the Cournot price-cost margins in equation (1.18).

Comparison with Ballester, Calvó-Armenagol and Zenou (2006). As I did for the Bertrand case, it is interesting to compare the result in Proposition 4 to the more general result provided in Ballester et al. Differently from Bertrand, the linear component of the quadratic payoff in (1.17) does not depend on the network structure, which makes the Cournot game analogous to the network game considered in Ballester et al., provided one defines the adjacency matrix as $-XX'$.

1.4.3 Bertrand vs Cournot

In this subsection, I compare the Bertrand and Cournot equilibrium outcomes. I start by describing the main differences between the two in terms of the product characteristics space. Then, I study how the network structure depends on the nature of competition and show that when there is strategic complementarity in prices, Bertrand competition leads to a network in which each product is more central and where price-cost margins are lower.

Product characteristics. In both models, product characteristics affect substitution patterns through an inner product matrix. In Cournot, this matrix is simply the inner product between each product's vector of attributes and is given by XX'. In Bertrand, as shown in Proposition 1, this inner-product matrix is instead $X\Omega^{-1}X'$ where Ω is a $K \times K$ matrix that reweights each vector of characteristics. At first glance, it might seem that product characteristics are different in the two models, but because Ω is positive definite, this turns out not to be the case.

To see this more formally, note that Ω is diagonalizable through an orthonormal basis of eigenvectors S such that $\Omega^{-1} = S\Lambda S'$ where $\Lambda = diag(\lambda_1, ..., \lambda_K)$ is a diagonal matrix that collects the eigenvalues of Ω^{-1}. Then, by letting $\tilde{X} = XS$, we can redefine product characteristics without changing their inner-product $\tilde{X}\tilde{X}' = XX'$. Thus comparing XX' to $X\Omega^{-1}X'$ is equivalent to comparing $\tilde{X}\tilde{X}'$ to $\tilde{X}\Lambda\tilde{X}'$. From this second comparison, it is easy to see that for any characteristic k, if $\tilde{x}_{jk} \geq \tilde{x}_{lk}$, then $\sqrt{\lambda_k}\tilde{x}_{jk} \geq \sqrt{\lambda_k}\tilde{x}_{lk}$, where the latter inequality holds because Ω is positive definite and thus $\lambda_k > 0$ for all k. In practice, this means that, under Bertrand, each product characteristic is just rescaled by a positive number λ_k. Thus, whether product j offers more characteristic k than product l is independent of the type of competition model, as one would expect.

Another important difference is that the competitive network implied by Bertrand does not depend on the units with which product characteristics are measured. From Proposition 3, we know that θ_{jl}, or equivalently $x_j'\Omega^{-1}x_l$, or equivalently $\tilde{x}_j'\Lambda\tilde{x}_l$, always lie in between -1 and 1 regardless of the units with which each characteristic is measured, which is an appealing property in an empirical context.

Competition and network structure. The following proposition shows how the nature of competition influences the structure of the competitive network by comparing the network centralities implied by price and quantity competition, respectively.

Proposition 5. *Assume that $\theta_{jk} \geq 0$ if $j \neq k$. Then each node centrality in the network*

implied by Bertrand competition is higher than the one under Cournot competition, i.e., $\tilde{\mathbf{b}}_b \geq -\tilde{\mathbf{b}}_c$ *where* $\tilde{\mathbf{b}}_b \equiv (I_J - diag(\Theta))^{-1/2}\mathbf{b}_b$ *and* $-\tilde{\mathbf{b}}_c \equiv (I_J - diag(\Theta^-))^{1/2}(-\mathbf{b}_c)$ *are the vectors of Bertrand and Cournot centralities respectively.*

As described in Section 1.2, several papers in the industrial organization literature have studied how equilibrium outcomes compare across the two types of competition models. In general, under strategic complementarity in prices, it is known that Bertrand competition with differentiated products is more efficient than Cournot competition and leads to lower prices and higher consumer surplus.[17] The same is true in this network setting where the vector of Bonacich centralities captures the intensity of competition. Proposition 5 shows that each node has a higher centrality under Bertrand and thus faces more intense competition. Moreover, looking back at the results in Propositions 3 and 4, one can see that the vector of Bonacich centralities is the only determinant of the wedge between Cournot and Bertrand prices. Hence, the effect of more intense competition on equilibrium prices must be entirely captured by differences in centrality.

Remarks. Before concluding this section, a couple of remarks are in order. First, note that the centrality implied by the Cournot model enters the inequality in Proposition 5 with a negative sign. This is because the centrality in Cournot is based on the matrix $-XX'$, which assigns a lower centrality to a more substitutable product. The measure $\mathbf{b}_b$ implied by Bertrand instead assigns higher centrality to more substitutable products and should be compared to $-\mathbf{b}_c$. To see this mathematically, suppose we knew that Cournot equilibrium prices p_c were to be higher than Bertrand prices p_b then by combining equations (1.13) and (1.18), the inequality in Proposition 5 would follow immediately.

Second, the Bonancich centralities are scaled by two positive diagonal matrices. This scaling appears because I am not imposing homogeneity across j of the θ_{jj} and θ^-_{jj} and

17. With homogenous products, the efficiency of Bertrand competition is maximized; even with only two firms, the perfectly competitive outcome obtains.

the vectors of Bonacich centralities enter equilibrium price-cost margin after this positive rescaling.[18] Furthermore, because the scaling is positive, the effect of product differentiation on equilibrium price-cost margins remains unaffected; higher centrality implies lower markups.[19]

1.5 Multiproduct Firms

In this section, I allow for the possibility that the same firm owns multiple products. I index firms by f and denote by J_f the set of products offered by firm f. Assuming there are $F < J$ firms, the $F \times J$ matrix R keeps track of which product belongs to which firm, i.e., $r_{fj} = \mathbf{1}\{j \in J_f\}$ and the $J \times J$ matrix $H = R'R$ denotes the ownership matrix. Moreover, as typically assumed in the context of multiproduct firms, I will assume that one product can only be owned by one firm or equivalently that the sets $(J_f)_{f=1}^F$ forms a partition of the set $\{1, ..., J\}$. The latter implies that the ownership matrix H will be a block-diagonal matrix with F blocks where the matrix of ones $1_{|J_f|}1'_{|J_f|}$ corresponds to f-th block.

In what follows, I will focus on Bertrand competition, but everything can be restated in terms of Cournot competition. I start by defining the profit maximization problem for a generic firm f and frame it as a Network game. Next, I show that in the case of multiproduct firms, the implied adjacency matrix only keeps track of the competitive links between products across different firms but not between products within the same firm. Finally, I extend the result in Proposition 3 and derive equilibrium price-cost margins in terms of the vector of Bonacich centralities. The result points to an intuitive comparative static for the

18. In Proposition 3 and 4 this homogeneity assumption was made only for expositional purposes. For more details, refer to the proofs of Propositions 3, 4 and 5, which are presented for the non-homogeneous case.

19. A similar type of rescaling also appears in the more general setting considered by Ballester, Calvó-Armenagol and Zenou (2006). If we allow for heterogeneity in the curvature (σ_{ii} in their notation) of own marginal returns, Remark 2 in their paper suggests scaling all elements of the adjacency matrix by its diagonal elements σ_{ii}. My scaling is similar, but it preserves the symmetry of the adjacency matrix.

conduct parameter H: when moving toward a more collusive industry structure (e.g., as $H \to 1_J 1_J'$), product differentiation (or equivalently product centrality) matters less and less for the equilibrium price-cost margins.

1.5.1 Multiproduct Firm Problem

Under Bertrand competition, firm f chooses prices $(p_j)_{j \in J_f}$ taking prices of competitors firms as given to maximize

$$\max_{(p_j)_{j \in J_f}} \sum_{j=1}^{J} r_{fj}(p_j - c_j) q_j(p_j, p_{-j}) \tag{1.19}$$

$$\text{s.t.} \quad q_j(p_j, p_{-j}) = \left(a_j(1 - \theta_{jj}) - \sum_{l \neq j} \theta_{jl} a_l \right) - (1 - \theta_{jj}) p_j + \sum_{l \neq j} \theta_{jl} p_l$$

Problem (1.19) is a quadratic problem in the vector of prices chosen by firm f, $(p_j)_{j \in J_f}$ and as such can be framed as a network game. The main difference from the single product problem in (1.10) is that now firm f will price its products jointly and will internalize how increasing a given price impacts the market shares of each of its products, i.e., the so-called portfolio effect.

How does the portfolio effect due to multiproduct pricing affect the equilibrium outcomes? When interpreting oligopolistic competition as a network game, the answer is fully captured by the adjacency matrix of the competitive network, which, as I show formally in Proposition 6, is a function of the ownership structure H

$$A_H(\Theta) \equiv \Theta - H \odot \Theta \tag{1.20}$$

where $\odot$ is the Hadamard matrix product. The matrix Θ still captures the competitive links between products. However, in the presence of multiproduct firms, the links between

products within the same firms are set to zero by subtracting the block diagonal matrix $H \odot \Theta$. The resulting adjacency matrix $A_H(\Theta)$ keeps track of the competitive links only between products owned by different firms.

1.5.2 Equilibrium price-cost Margins for Multiproduct Firms

In the next Proposition, I show that the result in Proposition 3 extends to the case in which firms sell multiple products and price them jointly. The result is presented for the case where own-price elasticities are heterogeneous across products, e.g., θ_{jj} varies across j.

Proposition 6. *For a given ownership structure H, if an interior equilibrium $(p_j^*)_{j=1}^J$ of the Bertrand pricing game exists, then it is unique and is such that*

$$p^* - c = \frac{\alpha - c}{2} - (I - H \odot \Theta)^{-1/2} \mathbf{b}\left(G_H(\Theta), \frac{1}{2}, (I - H \odot \Theta)^{1/2} \frac{\alpha - c}{2}\right) \tag{1.21}$$

provided $\max_j |\lambda_j(G)| \leq 2$ *and where* $G_H(\Theta) \equiv (I - H \odot \Theta)^{-1/2} A_H(\Theta) (I - H \odot \Theta)^{-1/2}$ *is a weighted adjacency matrix with elements in* $(-1, 1)$.

Similarly to the single-product case, in the multiproduct case, price-cost margins can be decomposed into a monopolistic component $(\alpha - c)/2$ and into a product differentiation component proportional to the vector of Bonacich centralities $\mathbf{b}$. The fact that some subsets of products are priced jointly influences the network structure and, in turn, the centrality of each node. In particular, it affects the adjacency matrix by setting all the competitive links between products owned by the same firm to zero, i.e., any (j, l) element of the matrix $A_H(\Theta)$ is zero whenever $i, j \in J_f$ for some firm f and the same is true for the matrix $G_H(\Theta)$.

One slight difference between the single and multiproduct cases is that the matrix $G(\Theta)$ that enters the Bonacich centrality is not necessarily symmetric, although its elements still belong to $(-1, 1)$. Luckily, this is not too much of a problem because one advantage of the Bonacich centrality compared to other measures is that it can be applied regardless of

whether the network is directed or undirected.[20] This would not be the case if we were to use a measure of degree centrality, which, in the case of directed networks, needs to consider the direction of the link.

Comparative static with respect to H. The expression in equation (1.21) points to an insightful comparative static in terms of the ownership structure matrix or, equivalently, the conduct parameter H. Consider first the extreme cases in which either firms are single-products (equivalently no collusion), or there is a single monopolist (equivalently perfect collusion). In the non-collusive case, we can see that expression (1.21) collapses to the outcome described in Proposition 3 because $H = I_J$ and $H \odot \Theta = diag(\Theta)$. Conversely, under perfect collusion, $H = 1_J 1_J'$, $H \odot \Theta = \Theta$ and price-cost margins collapse to the monopolistic price-cost margins given by $(\alpha - c)/2$. The reason is that when all products are priced jointly, all the competitive links in the adjacency matrix $A_{1_J 1_J'}(\Theta) = \Theta - \Theta = O$ are set to zero, and product differentiation does not matter for price-cost margins.

To gain more intuition, suppose that $A(\Theta)$ is non-negative or equivalently that prices are strategic complements. Then, as the market becomes more collusive (i.e., $H \to 1_J 1_J'$), the centrality of each product decreases because $A_H(\Theta) \to O$ element-wise and the importance of the monopolistic component (i.e., portfolio effect) in determining price-cost margins increases relative to the product differentiation component.

1.5.3 Market Definition and Mergers

Framing oligopolistic competition with product differentiation as a network game is insightful for at least two reasons.

First, from Proposition 6, we know that changes in the ownership structure will affect equilibrium price-cost margins only through the product differentiation component. Therefore, assuming there are no cost synergies, the only thing needed to assess the effect of

20. See also Remark 3 in Ballester, Calvó-Armenagol and Zenou (2006).

mergers on equilibrium prices is the vector of product centralities under the merger and no merger scenarios. In practice, this requires taking a stand on the relevant product characteristics for consumers. Depending on the industry under consideration, this might be more or less difficult. However, if one is willing to assume Bertrand competition, no normalization on the levels or the scale of product attributes is needed because, as I showed in Propositions 1 and 3, characteristics affect substitution patterns and price-cost margins only through a normalized inner-product.

A second advantage of modelling product differentiation as a competitive network is that, as already pointed out in Hoberg and Phillips (2016), there is no need to put too much thought into defining the relevant market. To the extent that the vector of attributes adequately describes substitution patterns across products, the implied network structure defines the market; products that are more substitutable to each other will share a stronger link in the network. Thus, even a broad definition of the market, such as an entire industry, would not affect the merger analysis as long as the product characteristics considered capture the true substitution patterns across products.

1.6 Application: The US Automobile Industry

In this section, I estimate the model using data on the US automobile industry. The same data have been used extensively in the empirical industrial organization literature starting from the seminal contribution in Berry, Levinsohn and Pakes (1995).

Before going into the estimation details, I want to remark that, although the car industry is important, it does not represent the most ideal application for the current setting. The reason is that consumers in the model have a taste for variety and will consume more than one good. In the context of car choice, this is not the most realistic assumption because households typically own no more than two cars and, in most cases, would buy only one car at the time of purchase. Nonetheless, as I show later, the model produces reasonable

substitution patterns and price-cost margins.

In what follows, I describe the data sources and the empirical model. Then, I recover own-cross price elasticities and price-cost margins. Lastly, I decompose price-cost margins and quantify how much of those are attributable to product differentiation.

1.6.1 Data

I obtain data on the US automobile industry from two different sources. First, I downloaded the data in Berry, Levinsohn and Pakes (1995) from the replication package accompanying a recent paper by Andrews, Gentzkow and Shapiro (2017). These data include the quantity sold by each car brand which I need because aggregate demand is derived in terms of quantities and not market shares. The second source of data is included in the recent Python package developed in Conlon and Gortmaker (2020), which again contains the very same automobile data but also includes the set of demand and supply instruments used in Berry, Levinsohn and Pakes (1995), which I will use to estimate demand in my model. Lastly, I collect data on the number of US households from the FRED website.

Overall, the data contains information on prices, quantities, market shares and characteristics of several car models sold in the US from 1971 to 1990. Table 1.1 reports sales-weighted averages of some relevant variables for each year. Average quantities sold are in units of 1000, average prices are in $1000 units, and the number of households is in millions. The last five columns are sales-weighted averages of product characteristics: HP/WT is the ratio of horsepower to weight, Air is a dummy for whether air conditioning is standard, Size captures the space of the car, MP$ measures the number of ten-mile increments one could drive for $1 worth of gasoline, and lastly MPG measures the number of ten-mile increments one could drive with one gallon of gasoline.

Several interesting patterns emerge from Table 1.1. The number of competing firms has been roughly constant, ranging between 17 and 22 multiproduct car manufacturers. The

Year	Firms	Models	Own models	Competitors	Quantity	Price	Households	HP/WT	Air	Size	MP$	MPG
1971	18	92	5.11	86.89	86.89	7.87	64.78	0.49	0.00	1.50	1.85	1.66
1972	19	89	4.68	84.32	98.62	7.98	66.68	0.39	0.01	1.51	1.87	1.62
1973	17	86	5.06	80.94	92.79	7.53	68.25	0.36	0.02	1.53	1.82	1.59
1974	17	72	4.24	67.76	105.12	7.51	69.86	0.35	0.03	1.51	1.45	1.57
1975	19	93	4.89	88.11	84.77	7.82	71.12	0.34	0.05	1.48	1.50	1.58
1976	21	99	4.71	94.29	93.38	7.79	72.87	0.34	0.06	1.51	1.70	1.76
1977	18	95	5.28	89.72	97.73	7.65	74.14	0.34	0.03	1.47	1.83	1.95
1978	18	95	5.28	89.72	99.44	7.64	76.03	0.35	0.03	1.40	1.93	1.98
1979	18	102	5.67	96.33	82.74	7.60	77.33	0.35	0.05	1.34	1.66	2.06
1980	19	103	5.42	97.58	71.57	7.72	80.78	0.35	0.08	1.30	1.47	2.21
1981	19	116	6.11	109.89	62.03	8.35	82.37	0.35	0.09	1.29	1.56	2.36
1982	19	110	5.79	104.21	61.89	8.83	83.53	0.35	0.13	1.28	1.82	2.44
1983	18	115	6.39	108.61	67.88	8.82	83.92	0.35	0.13	1.28	2.09	2.60
1984	20	113	5.65	107.35	85.93	8.87	85.41	0.36	0.13	1.29	2.12	2.47
1985	20	136	6.80	129.20	78.14	8.94	86.79	0.37	0.14	1.26	2.02	2.26
1986	22	130	5.91	124.09	83.76	9.38	88.46	0.38	0.18	1.25	2.86	2.42
1987	21	143	6.81	136.19	67.67	9.97	89.48	0.39	0.23	1.25	2.79	2.33
1988	20	150	7.50	142.50	67.08	10.07	91.07	0.40	0.24	1.25	2.92	2.33
1989	21	147	7.00	140.00	62.91	10.32	92.83	0.41	0.29	1.26	2.81	2.31
1990	20	131	6.55	124.45	66.38	10.34	93.35	0.42	0.31	1.27	2.85	2.27

Table 1.1: Sales-weighted averages.

same is true for the average number of car models produced by a single manufacturer, which increased slightly from 5 to 6.5. On the other hand, the total number of available car models increased more than 40%, from 92 in 1971 to 131 in 1990. At the same time, the average number of models produced by competitor firms also increased by 40%, from 87 models in 1971 to 124 in 1990, suggesting that competition in the number of products increased for the average firm. Next, looking at the average quantity and prices, the former decreased with cyclical ups and downs over time. In contrast, sales-weighted prices increased throughout the '80s while being roughly constant in the '70s. Product characteristics have also changed. For instance, air-conditioning becomes a more common feature over time, and both tens of miles per dollar (MP$) and tens of miles per gallon (MPG) increased, suggesting that cars have become more efficient over time. At the same time, car size seems to have decreased, whereas horsepower has been roughly constant.

Before turning to the estimation of the model and recovering products' price-cost margins, it is helpful to get a sense of how prices correlate with product characteristics in the raw

data without imposing any modelling structure. To this end, Figure 1.1 presents a binscatter of car prices against an unweighted measure of product centrality which summarizes how differentiated a product is relative to its competitors. The measure of centrality on the x-axis is motivated by the result I provided in Proposition 3 but differs in several respects because some terms are unobserved. The intuition described in equation (1.13) is simple: the margin over marginal cost a firm can charge depends negatively on how central its product is in the competitive network. Empirically though, whether or not this relationship between margins and centrality holds cannot be tested directly because marginal costs (c) and consumer preferences (α) are unobserved. Nonetheless, Figure 1.1 attempts to do so naively by proxying margins (i.e., the left-hand side of equation (1.13)) with observed prices and the Bonacich centrality $\mathbf{b}(A(\Theta), 1/2, (\alpha - c)/2)$ (i.e., the second term in (1.13)) with a similar centrality measure where $((\alpha - c)/2)$ is replaced with a vector of ones 1_J. The resulting measure is an unweighted Bonacich centrality given by

$$\mathbf{b}\left(A(\Theta), 1/2, 1_J\right) = \left(I_J - \frac{1}{2}A(\Theta)\right)^{-1} \frac{1}{2}A(\Theta)1_J, \tag{1.22}$$

which I can compute directly from the raw data. To compute Θ, I use the same product characteristics that enter the indirect utility in Berry, Levinsohn and Pakes (1995) namely horsepower (HP/WT), air conditioning, size and miles-per-dollar (MP$), and I calibrate η to 0.136.[21]

Returning to Figure 1.1, we can see that products with higher unweighted Bonacich centrality tend to charge lower prices. This decreasing relationship looks stronger for lower values of product centrality and seems to flatten as centrality increases. Although prices and centrality are negatively correlated as predicted by the network Bertrand model, from Figure 1.1, we cannot conclude that less central firms can charge higher margins and thus

21. The parameter η corresponds to the $\frac{\alpha}{1-\alpha}$ in Pellegrino (2023) which is calibrated to α to 0.12.

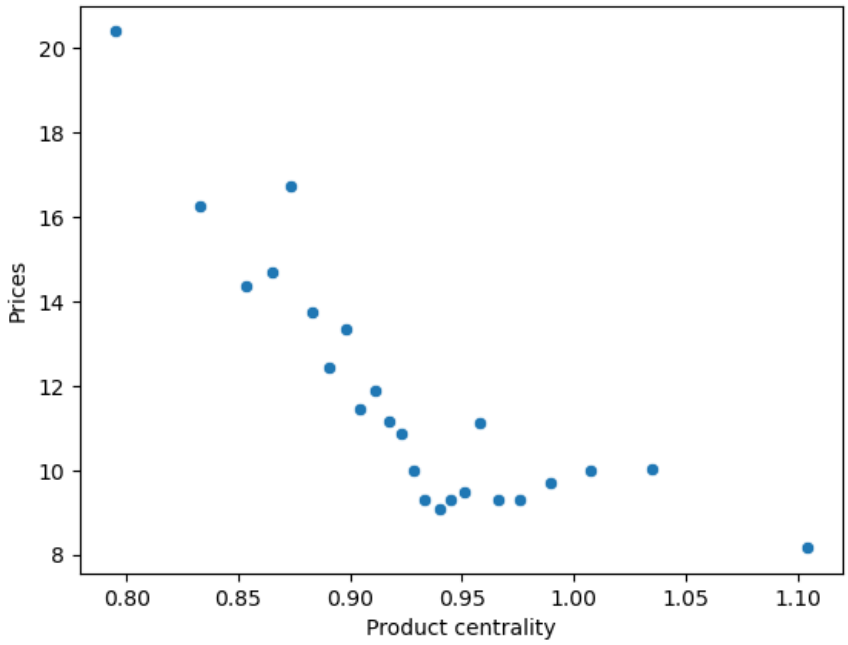

Figure 1.1: Binscatter of car prices against (unweighted) product Bonacich centrality after partialling out time fixed effects.

have more market power.[22] The reason is that the relationship can be driven by unobserved costs or heterogeneity in consumer preferences, which we are not accounting for. To uncover these unobservable components, we need a structural model of competition that allows us to identify consumer preferences and marginal costs. In the next section, I leverage the structure of the Bertrand network model to identify and estimate both the linear and quadratic components of consumer preferences. After estimating these demand parameters, I recover firms' marginal costs from the Nash-Bertrand equilibrium conditions.

1.6.2 The Empirical Nash-Bertrand Network Model

In this section, I make the Bertrand network model described in Section 1.4.1 empirically operational and adapt it to the context of car consumption.

An implicit assumption of the demand model introduced in Section 1.3 is the absence of income effects due to the quasi-linearity of consumer preferences in the outside good. In the context of car purchases, this assumption might be unrealistic. For this reason, I will now extend the demand model parsimoniously to accommodate income effects. To this end, I follow Berry, Levinsohn and Pakes (1995), and model consumer preferences for the inside and outside goods in a Cobb-Douglas fashion

$$U_i\left(y_i - p'q_i, q_i; X\right) = \left(y_i - p'q_i\right)^{\gamma} \left[G_i\left(q_i, X\right)\right]^{\phi} \tag{1.23}$$

where the first term in U already substitutes for the budget constraint and

$$G(q_i, X) \equiv \exp\left\{q_i'\alpha_i - \frac{\beta_i}{2} q_i'\left(I_J + \eta X'X\right) q_i\right\}.$$

22. Similarly, if prices and centrality were positively correlated or uncorrelated in the data, we could not conclude that centrality does not influence price-cost margins as predicted by the model.

Next, substituting G into (1.23) and taking logs, consumer i's utility can be written as

$$u_i(q_i, X) \equiv \log(U_i) = \gamma \log(y_i - p'q_i) + q_i'\alpha_i - \frac{\beta_i}{2} q_i' \left(I_J + \eta X'X\right) q_i \tag{1.24}$$

$$\approx q_i' \left(\alpha_i - \frac{p}{y_i}\right) - \frac{\beta_i}{2} q_i' \left(I_J + \eta X'X\right) q_i \tag{1.25}$$

where I normalize $\phi = 1$ and $\gamma = 1$ because they cannot be separately identified from α_i and β_i, and I use a first order Taylor expansion to approximate $\log\left(y_i - p'q_i\right)$.[23] The preferences in (1.25) are identical to the ones described in Section 1.3 except for the fact that prices are now measured relative to income. Consumer i's demand system can be derived as before

$$q_i(p) = \frac{1}{\beta_i} \left(I_J + \eta XX'\right)^{-1} \left(\alpha_i - \frac{p}{y_i}\right). \tag{1.26}$$

To derive the aggregate demand for product j we need to integrate over the distribution of consumer preferences (α_i, β_i, y_i). Assuming that y_i is independent of (α_i, β_i),[24] aggregate demand for product j is given by:

$$q_j(p_j, p_{-j}) = a_j - \frac{1}{\beta}(1 - \theta_{jj})\frac{p_j}{y} + \frac{1}{\beta} \sum_{k \neq j} \theta_{jk} \frac{p_k}{y} \tag{1.27}$$

where $\alpha = \int \frac{\alpha_i/\beta_i}{1/\beta_i} di$, $\beta = M \left[\int \frac{1}{\beta_i} di\right]^{-1}$, $y = M \left[\int \frac{1}{y_i} di\right]^{-1}$ and $\theta_{jk} = x_j' \Omega^{-1} x_k$. Overall, the demand functions are identical to the ones presented in Section 1.3 except that income effects are now present.

To make the model empirically operational, let t denote a market (i.e., a year in our

23. The same approximation has been used in Berry, Levinshon and Pakes (1999).

24. In empirical IO it commonly assumed that preferences parameters are idiosyncratic and independent from demographics.

empirical context) and note that equation (1.27) can be rearranged in vector form as

$$\tilde{q}_t \equiv (I_{J_t} + \eta X_t X_t')q_t = \frac{1}{\beta}\left(\alpha_t - \frac{p_t}{y_t}\right) \tag{1.28}$$

where J_t is the number of car models available in market t, X_t is a $J_t \times K$ matrix of product characteristics and the linear preference parameter vector α is allowed to vary over time. Next, consider the jth equation of the above system

$$\tilde{q}_{jt} = \frac{1}{\beta}\left(\alpha_{jt} - \frac{p_{jt}}{y_t}\right) \tag{1.29}$$

and note that, upon calibrating η and assuming that the matrix X_t contains only observables characteristics, the left-hand side in (1.27), denoted by $\tilde{q}_{jt}$, is directly measurable. Conversely, on the right-hand side of (1.27), only p_t and y_t are observable.[25]

More generally, equation (1.29) suggests that we can estimate demand using the following linear specification

$$\tilde{q}_{jt} = -\frac{1}{\beta}\frac{p_{jt}}{y_t} + w_{jt}'\zeta + \xi_{jt} \tag{1.30}$$

where w_{jt} is a vector of observable product and demographic characteristics which, in this context, includes both x_{jt} and y_t. On the other hand, ξ_{jt} includes characteristics that are unobservable to the econometrician but known by the agents.

To summarise, two assumptions allow us to estimate demand from (1.30). First, all the characteristics that enter consumer preferences in the quadratic term are observable. Second, any unobserved characteristic (ξ_{jt}) enters consumer preferences only through the

25. To measure income in a given year, I use the simulated draws that come with the pyBLP package developed in Conlon and Gortmaker (2020), and I average them using the weights provided. The draws come from a log-normal distribution of income whose location and scale parameters are estimated from the Current Population Survey (CPS) each year as described in Berry, Levinsohn and Pakes (1995).

linear parameter vector α, i.e.,[26]

$$\alpha_{jt} = \beta(w'_{jt}\zeta + \xi_{jt}) \tag{1.31}$$

To consistently estimate (1.30), we need to instrument p_{jt}/y_t because prices will be correlated with the unobservable component ξ_{jt}. The reason is that firms internalize ξ_{jt} before setting prices simultaneously. To see this formally, recall that Nash equilibrium prices are given by

$$p_{jt} = c_{jt} + \frac{y_t\alpha_{jt} - c_{jt}}{2} + \mathbf{b}_{jt} \tag{1.32}$$

and note that those prices are a function of α_{jt} (and in turn function of ξ_{jt}) both directly and indirectly through product j's Bonacich centrality $\mathbf{b}_{jt}$.[27] Under this setting, demand can be estimated with a simple linear instrumental variable strategy which I describe next.

1.6.3 Estimation Results

I start by estimating demand from the linear specification in (1.30). To do so, I instrument the term p_{jt}/y_t using the set of demand instruments z_{jt} constructed in Berry, Levinsohn and Pakes (1995) and available in the pyBLP Python package developed by Conlon and Gortmaker (2020). These instruments are a function of product characteristics and, as a consequence, are correlated with prices because, per the supply equation (1.32), characteristics affect firms' pricing decisions through the centrality term $\mathbf{b}_{jt}$ and the observable part of α_{jt}. The fact that our instruments are correlated with prices is not enough, and we also need to ensure that demand remains constant while instruments shift the supply. The identifying

26. I am also assuming that the parameter β is constant across markets. This assumption can be partially relaxed by interacting prices with any market level observable in equation (1.30).

27. The derivation of (1.30) is analogous to the one presented in the proof of Proposition (3) with the exception that the income term now appears.

	OLS	IV-2SLS
Constant	-0.1023	-0.1005
	(0.0043)	(0.0044)
Air (dummy)	-0.0179	0.0039
	(0.0145)	(0.0183)
MP$	0.0331	0.0368
	(0.0104)	(0.0104)
HP/WT	0.1432	0.1040
	(0.0448)	(0.0476)
Space	-0.0138	-0.0115
	(0.0168)	(0.0166)
price/income	-0.0043	-0.0327
	(0.0037)	(0.0121)
Fstat (Excluded)	-	92.1276
R2	0.8704	0.6257
Observations	2,217	2,217

Table 1.2: Demand estimates. Both specifications include interactions between characteristics and income, not reported here but available in Appendix 1.11. Standard errors are clustered at the car model level.

assumption relies on the idea that firms choose characteristics before observing any demand shock ξ_{jt} which formally boils down to requiring that $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$.

Table 1.2 reports both OLS and 2SLS demand estimates for the linear specification in equation (1.30). In both cases, the vector of characteristics w_{jt} includes a dummy for air conditioning, miles per dollar (MP$), horsepower (HP/WT), space and the interaction between all of those with income. The coefficients on characteristics are similar across the two specifications, and, in both cases, only horsepower and miles per dollar are significantly different from zero. Both coefficients are positive, suggesting that the average consumer prefers cars with more horsepower and cars that are more efficient in gasoline consumption. The estimated price-to-income coefficient is the most evident difference between the OLS and 2SLS specifications. With OLS, the coefficient is not statistically different from zero. In contrast, when we instrument for prices, the coefficient becomes negative and significant, and its magnitude increases almost ten folds in absolute value. The discrepancy between OLS

and IV estimates is quite common in contexts where prices and quantities are determined simultaneously in equilibrium. OLS estimates often imply inelastic demand curves because the observed variation in quantity and prices is also due to shifts in demand. However, after instrumenting for prices, the resulting estimates recover demand curves that are much more elastic.

After estimating the demand parameters, one would recover firms' marginal costs from equation (1.32). One issue in our context is that we do not observe α_{jt}, and we need to estimate it before being able to back out c_{jt}. Luckily, we can obtain an estimate of $\hat{\alpha}_{jt}$ for each product j and market t by simply plugging our demand estimates $(\hat{\beta}, \hat{\zeta})$ and the estimated regression residuals $(\hat{\xi}_{jt})$ into equation (1.31),

$$\hat{\alpha}_{jt} = \hat{\beta}\left(w'_{jt}\hat{\zeta} + \hat{\xi}_{jt}\right). \tag{1.33}$$

From the pricing equation in (1.32), with an estimate of α_{jt}, we can then recover marginal costs c_{jt}, price-cost margins $p_{jt} - c_{jt}$ and decompose the latter into monopolistic price-cost margins $(\alpha_{jt} - c_{jt})/2$ and network centrality $\mathbf{b}_{jt}$.

Figures 1.2 and 1.3 show the distribution of both marginal costs and price-cost margins in \$1000.[28] The median marginal cost across models and years is around \$6,900, and 3/4 of car models have a marginal cost lower than \$12,500. The distribution of estimated price-cost margin is more spread out, with a median margin around \$1,500 and 75% of car models charging a margin lower than \$3,200. Overall, although I started from linear-quadratic preferences rather than discrete choice, the estimated price-cost margins seem reasonably close to the ones recovered in Berry, Levinsohn and Pakes (1995).

Next, I decompose price-cost margins into a monopolistic component and a product centrality component which measures how differentiated a given product is relative to its com-

28. I removed roughly 6% of observations estimated to have negative marginal costs.

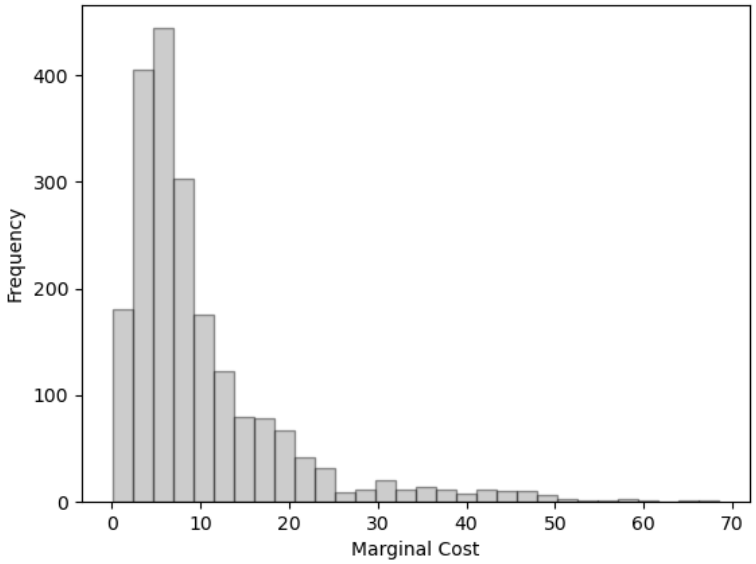

Figure 1.2: Distribution of marginal costs.

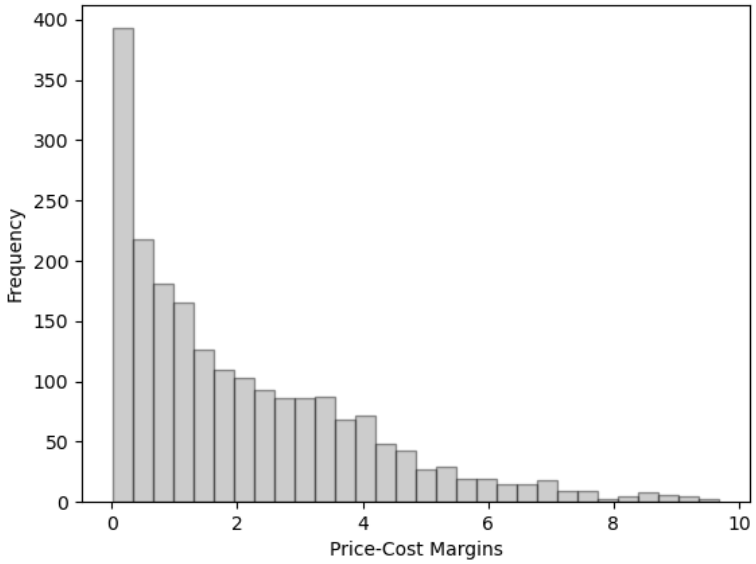

Figure 1.3: Distribution of price-cost margins.

Year	MC ($1000)	PMC ($1000)	MPCM ($1000)	BC ($1000)	PMC/MPCM (%)	BC/MPCM (%)
1971	4.435	4.668	92.937	88.269	4.99	95.01
1972	4.290	4.825	101.971	97.146	4.79	95.21
1973	4.096	4.695	89.996	85.301	5.26	94.74
1974	4.256	4.295	65.251	60.956	6.51	93.49
1975	4.462	3.803	66.107	62.304	5.70	94.30
1976	4.247	3.990	88.535	84.545	4.53	95.47
1977	4.299	4.014	93.829	89.815	4.33	95.67
1978	4.198	4.091	101.548	97.457	4.08	95.92
1979	4.541	3.708	70.299	66.592	5.24	94.76
1980	4.503	3.514	49.657	46.143	7.05	92.95
1981	5.627	3.151	49.460	46.309	6.36	93.64
1982	6.270	2.848	55.414	52.566	5.13	94.87
1983	5.906	3.281	75.745	72.464	4.39	95.61
1984	4.917	4.259	98.949	94.690	4.36	95.64
1985	5.767	3.614	101.260	97.645	3.62	96.38
1986	5.556	4.133	175.756	171.623	2.42	97.58
1987	6.578	3.807	150.768	146.960	2.56	97.44
1988	6.782	3.834	167.875	164.042	2.28	97.72
1989	7.288	3.600	146.416	142.816	2.46	97.54
1990	7.417	3.326	135.846	132.520	2.48	97.52

Table 1.3: Sales-weighted averages of the decomposition of price-cost margins (PMC).

petitors. Table 1.3 reports sales-weighted averages of such decomposition for each year available. The second column shows that the average marginal cost (MC) was stable throughout the '70s and then increased in the '80s. The third column reports average price-cost margins (PCM) in $1000. Following equation (1.13), the fourth and fifth columns decompose PCMs into monopolistic price-cost margins (MPCM) and a Bonacich centrality component (BC).[29] The monopolistic margins, given by $(\alpha_{jt} - c_{jt})/2$ vary considerably across years ranging from around $50,000 up to almost $170,000 and suggesting that the linear component of consumer preferences α_{jt} varies substantially over time. This time variation in preferences can capture changes in economic conditions, changes in regulation such as the import restriction on Japanese cars imposed throughout the '80s and changes in the type and number of car models available. The product centrality component $\mathbf{b}_{jt}$ also shows the same type of time variation and is, in magnitude, quite close to the monopolistic margins.

29. The decomposition used in Table 1.3 takes into account the fact that products have heterogeneous own-price elasticities and that firms are multiproduct.

Figure 1.4 takes a closer look by plotting the monopolistic margins (MPCM) together with the product centrality (BC) on the left y-axis and the price-cost margins on the right y-axis over time. Monopolistic margins and product centrality follow a similar cyclical behaviour suggesting that the main driver could be time variation in consumer preferences α_{jt}, which affects monopolistic margins but also the Bonacich product centrality through the weights.[30] Price-cost margins slightly fluctuate over time and seem to follow the cyclical behaviour of the other two variables. However, because the scale of the magnitude is way smaller, it is safe to conclude that PCMs have been basically constant over time, ranging on average between \$3000 and \$5000 overall but mostly in between \$3000 and \$4000 from 1975 onward.

The last two columns of Table 1.3 show what percentage of the monopolistic margins are captured by car manufacturers (PCM/MPCM) and what percentage of these margins is lost to competition between products (BC/MPCM). Not surprisingly, given the magnitudes observed in columns four and three, the last column shows that more than 90% of monopolistic price-cost margins are lost to competition, suggesting that the competitive network is dense and most of the products are close substitutes to each other. The reason for this could be that, although products are differentiated, a firm faces more than 85 competitors each year (i.e., Table 1.1 column four) on average, and each product likely has a close substitute in terms of the observable characteristics we are considering.[31] The second to last column shows the other side of the medal; firms only capture from 2% to 7% of the margins they could potentially charge in a monopolistic market.

Our estimates suggest that car manufacturers capture only a small fraction of the monopolistic margins they could charge. However, are those margins as tiny as they look? To answer this question, I compare the estimated PCMs (as a percentage of the MPCMs) with the margin a firm would charge in a Cournot game with homogeneous products. In

30. See equation (1.13) and Definition 1.

31. Recall that we are assuming that all characteristics that determine a product's centrality are observable to the econometrician.

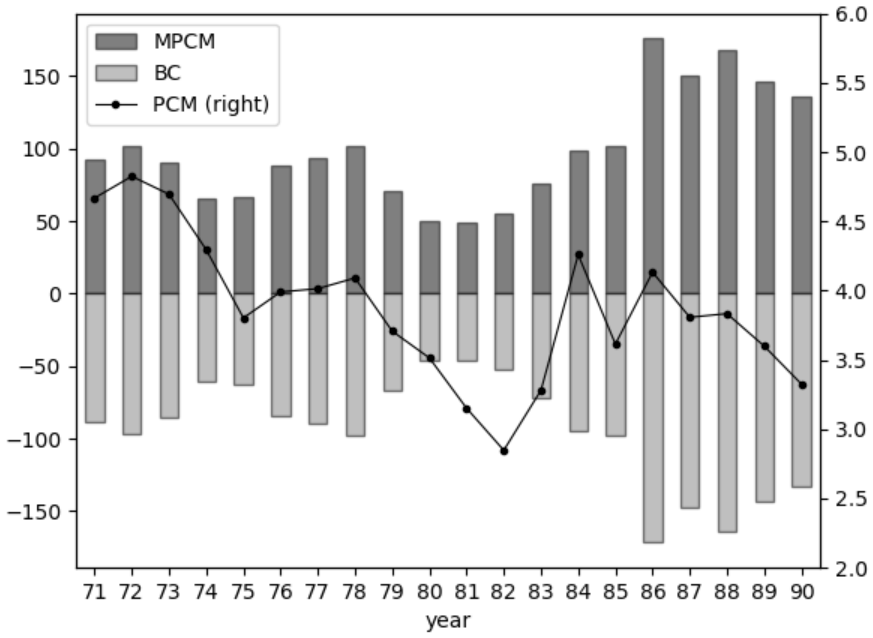

Figure 1.4: Decomposition of price-cost margins. All variables are measured in $1000

a homogenous Cournot model with N symmetric firms and linear demand, the equilibrium (dollar) price-cost margins are given by

$$p^{\text{hom. cournot}} - c = \frac{\alpha - c}{N+1} \tag{1.34}$$

where α is the demand intercept and c is the marginal cost. As a fraction of the monopolistic price-cost margin, the homogenous Cournot margins are only a function of the number of firms N

$$\frac{p^{\text{hom. cournot}} - c}{p^{\text{monopolist}} - c} = \frac{\alpha - c}{N+1} \cdot \frac{2}{\alpha - c} = \frac{2}{N+1}. \tag{1.35}$$

In Figure 1.5, I plot the estimated Bertrand PCM/MPCM reported in Table 1.3, together with the PCM/MPCM for the homogeneous Cournot model derived in equation (1.35) where

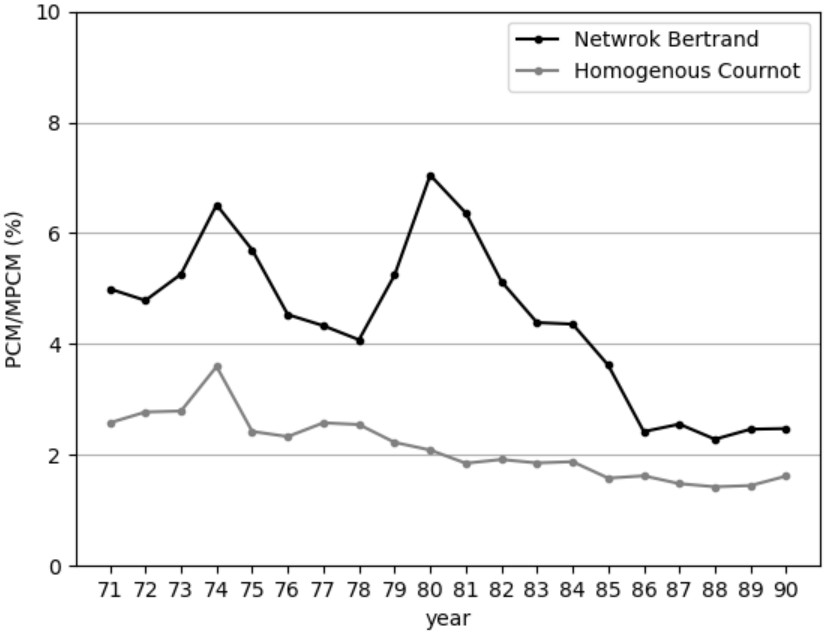

Figure 1.5: PCM/MPCM in Network Bertrand

I set the number of firms N equal to the average number of competitors a given car manufacturer faces in a given year (i.e., the number reported in Table 1.1) plus one. The pattern is clear; the margins firms charge in the Network Bertrand game are consistently above the margins under the homogeneous Cournot. Depending on the year, Bertrand margins can be more than three times higher than Cournot.[32] More generally, because the average number of competitors is the same across the two models, the difference between the two curves can be interpreted as the increase in margins that oligopolistic firms can capture when they offer differentiated products. Lastly, note that the difference between the Network Bertrand margins and the homogeneous Cournot margins represents a conservative estimate of the ability of firms to increase markups when products are differentiated. If we used a homogeneous Bertrand as benchmark, the increase would be more significant because firms, under

32. Interestingly, Bertrand margins peak throughout the '80s when a voluntary export restraint was placed on exports of automobiles from Japan to the United States, thereby reducing the supply of car models available.

homogeneous Bertrand, would be pricing at cost and charge zero margins.

I conclude the section by recovering own and cross-price elasticities implied by the demand estimates reported in Table 1.2. Recall from Section 1.3 that the (j, l) element of the elasticity matrix is given by

$$\epsilon_{jl} = \begin{cases} -\frac{1}{\beta}(1 - \theta_{jj}) & \text{if } j = l \\ \frac{1}{\beta}\theta_{jl} & \text{if } j \neq l \end{cases} \tag{1.36}$$

Because θ_{jl} is a function of observable product characteristics, we only need an estimate of $1/\beta$ to recover both own and cross-price elasticities.

Figure 1.6 plots a heatmap of the elasticity matrix for the year 1990. In that year, there were 131 car models available (Table 1.1), and Figure 1.6 shows the estimated own and cross-price elasticities for those models with an estimated own price elasticity larger than median.[33] Two remarks are in order. First, the magnitudes of own and cross-price elasticities are reasonable, with the own-price elasticities estimated to be negative and larger than the cross-price elasticities. Second, almost all cross-price elasticities are estimated to be small and positive. The latter would not be surprising in a model based on logit demand because that model restricts substitution patterns in a way that makes all products substitutes. However, this is not true in the linear-quadratic demand model I developed in Section 1.2. This type of quadratic preference does not impose any restrictions on product substitution patterns. However, the estimated cross-price elasticities are positive, suggesting that most car models are indeed substitutes for each other. Moreover, note that the fact that cross-price elasticities are positive is not a consequence of the fact that all product characteristics are positive numbers. More formally, having that $x_{jk} > 0$ for any product j and characteristic k does not imply that all products are substitutes or, equivalently, that θ_{jl} is always positive.

33. This is done for visibility purposes. Estimates remain reasonable if we look at all models with own-price elasticity above the 25th percentile, reported in Appendix 1.11 Figure 1.7

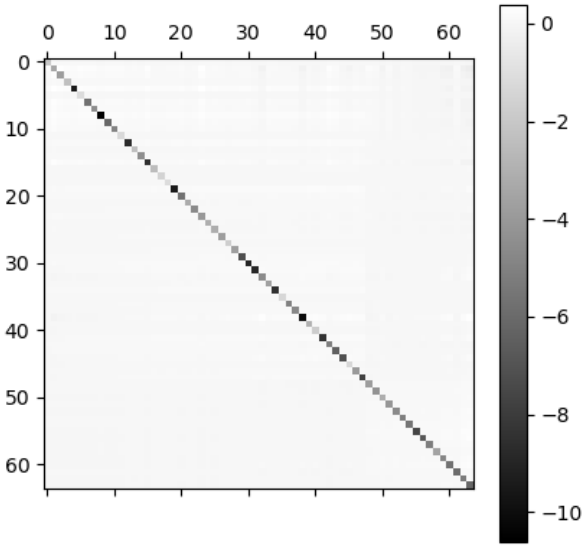

Figure 1.6: Matrix of estimated own and cross price elasticities for the year 1990. The products included are the ones with an estimated own price elasticity above the median.

1.7 Conclusion

This paper studies how product differentiation affects substitution patterns, and firms' price-cost margins in oligopolistic markets where products are differentiated over multiple attributes and consumers have linear-quadratic preferences. Under these assumptions, oligopolistic competition in either prices or quantities can be framed as a network game where a product location in the network is determined by its vector of attributes, and the network links between products capture the extent with which two products compete. Products with similar characteristics will be closer to each other and will compete more intensely. On the other hand, products with more unique characteristics will have a more peripheral location and enjoy more market power.

More precisely, I show that firms' price-cost margins can be decomposed additively into a monopolistic component and a product differentiation component. The first component coincides with the margins a monopolist would charge. In contrast, the second component

is proportional to how central a product is in the network as measured by its Bonacich centrality. Products with higher centrality must give up a more significant fraction of the monopolistic margins because they face more competition. Furthermore, I showed that this decomposition holds for both Bertrand and Cournot price-cost margins and can be extended to allow the presence of multiproduct firms.

In the second part of the paper, I show how to estimate the model using market-level data on a given industry. Under the assumption that unobserved product characteristics enter consumer preferences only through the linear component of the utility, a simple linear IV strategy identifies the demand parameters. Then, marginal costs and margins can be recovered from the Nash equilibrium pricing equations. In addition, using the decomposition of price-cost margins, one can quantify what fraction of the potential monopolistic margins a firm can capture by differentiating its product from its competitors.

In the last part of the paper, I estimate the model using data on the US automobile industry from 1971 to 1990. I find that it delivers substitution patterns and price-cost margins comparable to the ones estimated in the literature that models automobile demand starting from individual discrete choice problems. Interestingly, although the linear-quadratic demand model does not restrict substitution patterns in any way, the estimated cross-price elasticities are almost always positive, suggesting that cars are substitutes, something the discrete choice demand framework instead imposes a priori. Finally, I decompose firms' price-cost margins and find that car manufacturers capture from 2% to 7% of the monopolistic margins depending on the year. These margins can be as high as three times the margins firms would be able to charge if their products were to be homogeneous.

1.8 Appendix: Proofs

Proof of Proposition 1. Taking the horizontal sum of individual demands in (1.3) it is easy to check that the aggregate demand system is given by

$$q(p) = \int q_i(p)\frac{di}{M} \tag{1.37}$$

$$= \frac{1}{\eta\beta}\left(\frac{1}{\eta}I_J + XX'\right)^{-1}(\alpha - p) \tag{1.38}$$

where $\beta \equiv \int \frac{1}{\beta_i}\frac{di}{M}$ and $\alpha \equiv \int \frac{\alpha_i/\beta_i di}{\int 1/\beta_i di}$. Next, note that

$$q(p) = \frac{1}{\eta\beta}\left(\frac{1}{\eta}I_J + XX'\right)^{-1}(\alpha - p) \tag{1.39}$$

$$= \frac{\alpha - p}{\beta} - X\Omega^{-1}X'\frac{(\alpha - p)}{\beta} \tag{1.40}$$

where $\Omega \equiv \frac{1}{\eta}I_K + X'X$. Then let $a \equiv \frac{\alpha}{\beta} - X\Omega^{-1}X'\frac{\alpha}{\beta}$ and $\Theta \equiv X\Omega^{-1}X'$ to obtain the expression in the main text. To complete the proof, we are left to show that $\theta_{jj} \in (0,1)$ and that $\theta_{jl} \in (-1,1)$ for any $j \neq l$. To this end, first note that by construction both Θ and $I_J - \Theta$ are positive definite matrices. Then, letting e_j be the j-th unit vector, we have that $\theta_{jj} = e_j'\Theta e_j > 0$ and similarly $1 - \theta_{jj} = e_j'(I - \Theta)e_j > 0$. Next, take any (j,l) pair with $j \neq l$ and note that

$$\theta_{jj} + \theta_{ll} - 2\theta_{jl} = (e_j - e_l)'\Theta(e_j - e_l) > 0 \tag{1.41}$$

where the first equality exploits the fact that Θ is symmetric. From (2.40) and the fact that $\theta_{jj} < 1$ for all j, we can conclude that $\theta_{jl} < 1$. To show that $\theta_{jl} > -1$ it is enough to repeat the previous argument using $\theta_{jj} + \theta_{ll} + 2\theta_{jl}$.

Proof of Proposition 2 Because U is the matrix of principal component directions of X, it is also the matrix whose columns correspond to eigenvectors of $X'X$. Then we have,

assuming without loss that $\eta = 1$,

$$\Theta = X(I_K + X'X)^{-1}X' \tag{1.42}$$

$$= X(I_K + U\Lambda^{x'x}U')^{-1}X' \tag{1.43}$$

$$= XU(I_K + \Lambda^{x'x})^{-1}U'X' \tag{1.44}$$

$$= \tilde{X}(I_K + \Lambda^{x'x})^{-1}\tilde{X}' \tag{1.45}$$

where $\Lambda^{x'x}$ is the diagonal matrix that collects the eigenvalues of $X'X$. For $\eta \neq 1$, it is enough to redefine $X = \sqrt{\eta}X$ to obtain the same result.

Proof of Proposition 3 The proof presented is for the more general case in which θ_{jj} are heterogeneous across j. The first order condition of (1.10) with respect to p_j reads:

$$\alpha_j(1-\theta_{jj}) - \sum_{l\neq j}\theta_{jl}\alpha_l + c_j(1-\theta_{jj}) - 2(1-\theta_{jj})p_j + \sum_{l\neq j}\theta_{jl}p_l = 0 \tag{1.46}$$

after rearranging in vector form and solving for p^* one obtains

$$\begin{aligned} p^* - c &= \frac{1}{2}\left(I_J - diag(\Theta) - \frac{A(\Theta)}{2}\right)^{-1}(I-\Theta)(\alpha - c) \\ &= \frac{1}{2}\left(I_J - \left(I_J - diag(\Theta) - \frac{A(\Theta)}{2}\right)^{-1}\frac{A(\Theta)}{2}\right)(\alpha - c) \\ &= \frac{\alpha - c}{2} - (I_J - diag(\Theta))^{-1/2}\left(I_J - \frac{G(\Theta)}{2}\right)^{-1}\frac{G(\Theta)}{2}(I - diag(\Theta))^{1/2}\frac{(\alpha - c)}{2} \\ &= \frac{\alpha - c}{2} - (I_J - diag(\Theta))^{-1/2}\,\mathbf{b}\left(G(\Theta), \frac{1}{2}, (I - diag(\Theta))^{1/2}\frac{\alpha - c}{2}\right) \end{aligned} \tag{1.47}$$

where

$$G(\Theta) \equiv (I - diag(\Theta))^{-1/2}\,A(\Theta)\,(I - diag(\Theta))^{-1/2} \tag{1.48}$$

and the last equality is well-defined provided $\max_j |\lambda_j(G)| < 2$. Expression (1.13) then obtains when imposing $\theta_{jj} \equiv \theta$. To complete the proof for the more general case, I need to show that $G(\Theta)$ preserves the properties of $A(\Theta)$. It is immediate to see that $G(\Theta)$ is 0-diagonal and symmetric. I am left to show that all its off-diagonal elements g_{ij} lie in $(-1, 1)$. To see this recall that $I_J - \Theta$ is positive definite and note that

$$(I_J - diag(\Theta))^{1/2}(I_J - G(\Theta))(I_J - diag(\Theta))^{1/2} = I_J - \Theta \tag{1.49}$$

which implies that

$$1 - g_{ij}^2 = (g_{ij}e_i + e_j)'(I_J - G(\Theta))(g_{ij}e_i + e_j) > 0 \tag{1.50}$$

which completes the proof.

Proof of Proposition 4 The proof presented is for the more general case in which θ_{jj}^- are heterogeneous across j. To find Cournot price-cost margins we first need to find the Cournot equilibrium quantities which are given by

$$q^* = \left(2I_J - \eta diag(\Theta^-) - \eta\Theta^-\right)^{-1} \frac{\alpha - c}{\beta} \tag{1.51}$$

where $\Theta^- \equiv -XX'$. Plugging (1.51) into the aggregate inverse demand one obtains

$$\begin{aligned}
p^* - c &= \alpha - c - \beta \left(I_J - \eta\Theta^-\right) q^* \\
&= \frac{\alpha - c}{2} + \eta\frac{A(\Theta^-)}{2}\left(I_J - \eta diag(\Theta^-) - \eta\frac{A(\Theta^-)}{2}\right)^{-1}\frac{\alpha - c}{2} \\
&= \frac{\alpha - c}{2} + \left(I_J - \eta diag(\Theta^-)\right)^{1/2}\left(\eta\frac{G(\Theta^-)}{2}\right)\left(I_J - \eta\frac{G(\Theta^-)}{2}\right)^{-1} \times \\
&\times \left(I_J - \eta diag(\Theta^-)\right)^{-1/2}\frac{\alpha - c}{2} \\
&= \frac{\alpha - c}{2} + \left(I_J - \eta diag(\Theta^-)\right)^{1/2} \times \\
&\times \mathbf{b}\left(G(\Theta^-), \frac{\eta}{2}, \left(I_J - \eta diag(\Theta^-)\right)^{-1/2}\frac{\alpha - c}{2}\right)
\end{aligned} \tag{1.52}$$

where

$$G(\Theta^-) \equiv \left(I_J - diag(\Theta^-)\right)^{-1/2} A(\Theta^-) \left(I_J - diag(\Theta^-)\right)^{-1/2} \tag{1.53}$$

where the last equality is well-defined provided $\eta < \max_j |\lambda(G)|/2$. Expression (1.18) in the main text then obtains by replacing θ^-_{jj} with θ^- for any j and thus completes the proof. Note that, as mentioned in the main text, without any normalization the elements of the matrix Θ^- can take any value possibly outside $(-1, 1)$. While this is not too much of a problem mathematically, it is unappealing conceptually because it implies that the adjacency matrix $A(\Theta^-)$ can have weights greater than 1 in absolute value.

Proof of Proposition 5 Let f and g denote the inverse demand and demand systems respectively:

$$p = f(q) = \alpha - \beta\left(I_J + \eta XX'\right) q \tag{1.54}$$

$$q = g(p) = (I_J - \Theta)\frac{(\alpha - p)}{\beta} \tag{1.55}$$

where $\Theta = X\Omega^{-1}X'$ and $\Omega = \frac{1}{\eta}I_K + X'X$. Firm j's first order condition of the Bertrand problem can be written as

$$g_j(p) + (f_j(g(p)) - c)\frac{\partial g_j(p)}{\partial p_j} \tag{1.56}$$

$$= g_j(p)\sum_{l\neq j}\frac{\partial f_j}{\partial g_k}\frac{\partial g_k}{\partial p_j} + \left(f_j(g(p)) - c + \frac{\partial f_j}{\partial g_j}\right)\frac{\partial g_j(p)}{\partial p_j} \tag{1.57}$$

where the second equality differentiates the identity $p_j \equiv f_j(g(p))$ with respect to p_j. Next, denote the Cournot equilibrium price p^c and note that the above Bertrand FOC evaluated at p^c reduces to

$$g_j(p^c)\sum_{l\neq j}\frac{\partial f_j}{\partial g_k}\frac{\partial g_k}{\partial p_j} = g_j(p^c)\sum_{l\neq j}(-\beta\eta x_j' x_k)\frac{\theta_{kj}}{\beta} < 0 \tag{1.58}$$

where the last inequality holds because $\theta_{kj} \geq 0$ implies that $x_j' x_k \geq 0$. To see this note that both $I_J - \Theta$ and Θ are positive definite matrices. But this implies that all eigenvalues of Θ must lie in $(0, 1)$ and, consequently

$$(I_J + \eta XX') = \tag{1.59}$$

$$= \left(I_J - X\left(\frac{1}{\eta}I_K + X'X\right)^{-1}X'\right)^{-1} \tag{1.60}$$

$$= (I_J - \Theta)^{-1} = \sum_{s=1}^{\infty}\Theta^s \geq 0, \tag{1.61}$$

where the last equality is well defined because $\max_j |\lambda_j(\Theta)| < 1$. Thus, because Θ is non-negative by assumption we can conclude that $x_j' x_k \geq 0$ if $j \neq k$. Because equation (1.58) holds for all j we can consider the vector of Bertrand FOCs evaluated at the Cournot price

p_c

$$(I_J - \Theta)\alpha + (I_J - diag(\Theta))c - 2\left(I_J - \frac{diag(\Theta)}{2} - \frac{\Theta}{2}\right)p_c < 0 \tag{1.62}$$

$$\Leftrightarrow p_b = \frac{1}{2}\left(I_J - \frac{diag(\Theta)}{2} - \frac{\Theta}{2}\right)^{-1}((I_J - \Theta)\alpha + (I_J - diag(\Theta))c) < p_c \tag{1.63}$$

where p_b denotes the Bertrand equilibrium price vector. Then, from Propositions 3 and 4 we can conclude that Bertrand centralities are higher than the ones implied by Cournot:

$$\tilde{\mathbf{b}}_b \equiv (I_J - diag(\Theta))^{-1/2}\,\mathbf{b}_b \geq \left(I_J - \eta diag(\Theta^-)\right)^{1/2}(-\mathbf{b}_c) \equiv -\tilde{\mathbf{b}}_c. \tag{1.64}$$

Proof of Proposition 6 The proof is identical to the proof of Proposition 3 with the difference that the matrix $diag(\Theta)$ must be replaced by $H \odot \Theta$. The first order condition of (1.19) with respect to p_j is given by

$$a_j + c_j - \sum_{l \in J_f} \theta_{lj} c_l - 2(1 - \theta_{jj})p_j + \sum_{l \in J_f/\{j\}} \theta_{lj} p_l + \sum_{l \neq j} \theta_{jl} p_l = 0 \tag{1.65}$$

and after rearranging in vector form and solving for p^* one obtains

$$p^* - c = \frac{1}{2}\left(I_J - H \odot \Theta - \frac{A_H(\Theta)}{2}\right)^{-1}(I_J - \Theta)(\alpha - c) \tag{1.66}$$

$$= \frac{1}{2}\left[I_J - \left(I_J - H \odot \Theta - \frac{A_H(\Theta)}{2}\right)^{-1}\frac{A_H(\Theta)}{2}\right](\alpha - c) \tag{1.67}$$

$$= \frac{\alpha - c}{2} - (I_J - H \odot \Theta)^{-1/2}\left(I_J - \frac{G_H(\Theta)}{2}\right)^{-1}\frac{G_H(\Theta)}{2} \times \tag{1.68}$$

$$\times (I_J - H \odot \Theta)^{1/2}\frac{(\alpha - c)}{2} \tag{1.69}$$

$$= \frac{\alpha - c}{2} - (I_J - H \odot \Theta)^{-1/2}\, b\left(G_H(\Theta), \frac{1}{2}, (I_J - H \odot \Theta)^{1/2}\frac{(\alpha - c)}{2}\right) \tag{1.70}$$

where

$$G_H(\Theta) \equiv (I_J - H \odot \Theta)^{-1/2} A_H(\Theta) (I_J - H \odot \Theta)^{-1/2} \tag{1.71}$$

and the last equality is well-defined provided $\max_j |\lambda_j(G)| < 2$, where λ_j is the an eigenvalue of $G_H(\Theta)$. To complete the proof I need to show that $G_H(\Theta)$ is a weighted adjacency matrix. First note that because $A_H(\Theta) = \Theta - H \odot \Theta$ is a 0-block diagonal matrix the same holds for $G_H(\Theta)$. Next, I show that all its elements g_{ij} lie in $(-1, 1)$. To see this recall that $I_J - \Theta$ is positive definite and note that

$$(I_J - H \odot \Theta)^{1/2} (I_J - G_H(\Theta)) (I_J - H \odot \Theta)^{1/2} = I_J - \Theta \tag{1.72}$$

which implies that $(I_J - G_H(\Theta))$ is also positive definite, but then

$$1 - g_{ij}^2 = (g_{ij} e_i + e_j)' (I_J - G_H(\Theta)) \left(g_{ij} e_i + e_j\right) > 0. \tag{1.73}$$

So far we have used extensively $(I_J - H \odot \Theta)^{1/2}$, to make sure that this is well defined I need to show that $I_J - H \odot \Theta$ is positive definite. To see this, note that $I_J - \Theta > 0$ implies that all its principal submatrices are positive definite, but then all blocks in the block-diagonal matrix $H \odot (I_J - \Theta)$ are also positive definite which makes $H \odot (I_J - \Theta)$ positive definite as well.

1.9 Equilibrium existence

Throughout the analysis I have assumed that an interior Nash equilibrium exists. In this appendix, I provide conditions for the existence and uniqueness of the equilibrium. Specifically, I show that the following dominance-diagonal type of condition,

$$(1-\theta_{jj})(\alpha_j - c_j) \geq \sum_{k\neq j} |\theta_{jk}|(\alpha_k - c_k) \quad \text{all } j \tag{1.74}$$

ensures that each firm j charges a price p_j that belongs to the interval $[c_j, \alpha_j]$ when competing a la Bertrand.[34] Recall, that the system of FOCs is linear in the vector of prices p and given by

$$(I-\Theta)p + (I - diag(\Theta))p = (I-\Theta)\alpha + (I - diag(\Theta))c.$$

The above system has a well-defined solution as long as $(I-\Theta)$ is non-singular, which is always the case because $(I-\Theta)$ is a positive definite matrix. Moreover, linearity would also imply that such solution is unique. What we do not know is whether the equilibrium prices are non-negative and above firms' marginal costs. The above dominance diagonal condition ensures that the system of best replies is a self-map over the set $\times_{j\in\{1,\ldots,J\}}[c_j, \alpha_j]$. To start with define the linear operator $T : \mathbb{R}^J \to \mathbb{R}^J$ whose jth component is fund j best reply:

$$T_j(p) = \frac{1}{2}\left[\alpha_j - c_j - \sum_{k\neq j} \frac{\theta_{jk}}{1-\theta_{jj}}(\alpha_k - p_k)\right].$$

34. Clearly, there is an implicit assumption here that consumers willingness to pay for the very first unit α_j (e.g., the demand intercept) is greater than the marginal cost c_j.

Next, we can show that whenever $p \in \times_{j\in\{1,...,J\}}[c_j, \alpha_j]$, condition (1.74) implies that $T_j \in [c_j, \alpha_j]$. To see this, take any $p \in \times_{j\in\{1,...,J\}}[c_j, \alpha_j]$ and note that

$$\begin{aligned} & c_j \leq T_j(p) \leq \alpha_j \\ \Leftrightarrow & \left| \sum_{k\neq j} \frac{\theta_{jk}}{1-\theta_{jj}} \frac{\alpha_k - c_k}{\alpha_j - c_j} \frac{\alpha_k - p_k}{\alpha_k - c_k} \right| \leq 1 \\ \Leftrightarrow & \sum_{k\neq j} \frac{|\theta_{jk}|}{1-\theta_{jj}} \frac{\alpha_k - c_k}{\alpha_j - c_j} \frac{\alpha_k - p_k}{\alpha_k - c_k} \leq 1 \end{aligned}$$

is always satisfied when (1.74). Because T maps a closed and bounded set into itself, Bower fixed point theorem implies that there exists an $p^* \in \times_{j\in\{1,...,J\}}[c_j, \mu_j]$ such that $T(p^*) = p^*$. Moreover, from (2.58) we know that such fixed point is unique, because T is linear.

1.10 Decomposition of Cournot Markups

In this section, I show how to decompose the markups of a given product j in the case of Cournot competition and formally show that the product centrality measure defined in Pellegrino (2023), and denoted by $1-\chi_j$, is an affine transformation of the Bonacich centrality measure $\mathbf{b}_j$ I have been using throughout this paper.

To start with, I make a few notational adjustments to conform my notation to the one in Pellegrino (2023). I normalize the squared norm of product characteristics to one, i.e., for any product j, I assume that $||x_j||_2^2 = x_j' x_j = 1$. Following the notation I used in Proposition 4, this normalization is equivalent to set $\theta^- = -1$. Also, I set $\beta_i \equiv 1$ for any consumer i, which further implies that in the aggregate demand derived in Proposition 1, $\beta = 1$ and the aggregate demand intercept $\alpha = M^{-1} \int \alpha_i di$. In matrix form, the aggregate demand and its

inverse are given by

$$q = \left(I + \eta XX'\right)^{-1}(\alpha - p) \tag{1.75}$$

$$p = \alpha - (I + \eta XX')q. \tag{1.76}$$

Next, let ν be such that $\eta = \frac{\nu}{1-\nu}$ and note that

$$I + \eta XX' = \frac{1}{1-\nu}I + \frac{\nu}{1-\nu}(XX' - I) \tag{1.77}$$

where the ν plays the role of α in the notation of Pellegrino (2023). Next, assuming that $(1-\nu) > 0$, I can re-scale consumer preferences in 1.1 by $(1-\nu)$ to obtain

$$v_i(q_i, X) \equiv (1-\nu)u_i(q_i, X) = q_i'(\alpha_i - p) - \frac{1}{2}q_i'(I + \Sigma)q_i \tag{1.78}$$

where $\Sigma \equiv \nu(XX' - I)$, I redefined $\alpha_i \equiv \nu X\alpha_i^x + (1-\nu)\alpha_i^q$ and I already substituted in for the budget constraint.[35] The preferences in (1.78) are exactly the same ones used in Pellegrino (2023) and the associated aggregate demand and its inverse are

$$q = (I + \Sigma)^{-1}(\alpha - p) \tag{1.79}$$

$$p = \alpha - (I + \Sigma)q. \tag{1.80}$$

Next, I show how the product centrality measure defined in Pellegrino (2023) and denoted by $1 - \chi_j$ is an affine transformation of product j's Bonacich network centrality $\mathbf{b}_j(-\Sigma, 1/2, (\alpha - c)/2)$. The steps are similar to the ones in Proposition 4 with $\Theta^- = -\Sigma$.

35. Note that it is not necessary to rescale the outside good q_{i0} by $(1-\nu)$ because I can always normalize its price to $1/(1-\nu)$ in the budget constraint.

To start with, note that the Cournot equilibrium quantities are given by

$$q^* = \left(1 - \frac{1}{2}(-\Sigma)\right)^{-1} \frac{\alpha - c}{2}. \tag{1.81}$$

Replacing (1.81) in the inverse demand, we get the price-cost margin decomposition obtained in Proposition (4)

$$p^* - c = \alpha - c + \left(I - \frac{1}{2}(-\Sigma)\right)^{-1} \frac{1}{2}(-\Sigma)\frac{\alpha - c}{2} \tag{1.82}$$

$$= \frac{\alpha - c}{2} + \mathbf{b}\left(-\Sigma, \frac{1}{2}, \frac{\alpha - c}{2}\right). \tag{1.83}$$

Next, define product j markups as

$$\mu_j \equiv \frac{p_j}{c_j} \tag{1.84}$$

and letting $\mathbf{b}_j$ the j-th component of the vector of Bonacich centralities $\mathbf{b}$, from (1.83) we obtain

$$\mu_j = \frac{\alpha_j + c_j}{2c_j} + \frac{1}{c_j}\mathbf{b}_j = \bar{\mu}_j + \frac{1}{c_j}\mathbf{b}_j \tag{1.85}$$

where $\bar{\mu}_j$ denotes the monopolistic markup.

Pellegrino (2023) shows that product j markup is given by

$$\mu_j = \chi_j + (1 - \chi_j)\bar{\mu_j} \tag{1.86}$$

where the product j's centrality $1 - \chi_j$ is defined as

$$1 - \chi_j \equiv \frac{1}{\alpha_j - c_j}\left[\left(I + \frac{1}{2}\Sigma\right)^{-1}(\alpha - c)\right]_j \tag{1.87}$$

where for a genric vector y, $[y]_j$ denotes its j-th component. Combining (1.85) and (1.86) and solving for χ_j one obtains

$$\chi_j = -\frac{2}{\alpha_j - c_j}\mathbf{b}_j \tag{1.88}$$

from which it is immediate to see that Pellegrino (2023)'s centrality measure is an affine transformation of the Bonacich centrality

$$1 - \chi_j = 1 + \frac{2}{\alpha_j - c_j}\mathbf{b}_j. \tag{1.89}$$

To complete the argument, I am left to show that from (1.89) I can back out the definition of product centrality in (1.87). To see this, let χ be the J-vector with j-th component χ_j and building from (1.89)

$$1_J - \chi = [diag(\alpha - c)]^{-1}(\alpha - c + 2\mathbf{b}) \tag{1.90}$$

$$= [diag(\alpha - c)]^{-1}\left[I - \left(I + \frac{1}{2}\Sigma\right)^{-1}\Sigma\frac{1}{2}\right](\alpha - c) \tag{1.91}$$

$$= [diag(\alpha - c)]^{-1}\left(I + \frac{1}{2}\Sigma\right)^{-1}(\alpha - c) \tag{1.92}$$

where 1_J is the J-vector of ones, the first equation is just (1.89) in vector forms, the second equation substitutes for the definition of Bonacich centrality and the last equation rearrange terms. Overall, the j-th equation in (1.92) coincides with the definition of product centrality for product j in (1.87), thus proving that the argument is consistent.

1.11 Appendix: Additional Figures and Tables

	OLS	IV-2SLS
Constant	-0.1023	-0.1005
	(0.0043)	(0.0044)
Air (dummy)	-0.0179	0.0039
	(0.0145)	(0.0183)
MP$	0.0331	0.0368
	(0.0104)	(0.0104)
HP/WT	0.1432	0.1040
	(0.0448)	(0.0476)
Space	-0.0138	-0.0115
	(0.0168)	(0.0166)
air×income	0.0001	0.0000
	0.0001	0.0001
MP$×income	0.0001	0.0001
	0.0001	0.0001
HP/WT×income	-0.0005	-0.0003
	0.0002	0.0003
Space×income	0.0004	0.0004
	0.0001	0.0001
price/income	-0.0043	-0.0327
	(0.0037)	(0.0121)
Fstat (Excluded)	-	92.1276
R2	0.8704	0.6257
Observations	2,217	2,217

Table 1.4: Demand estimates. Standard errors are clustered at the model level.

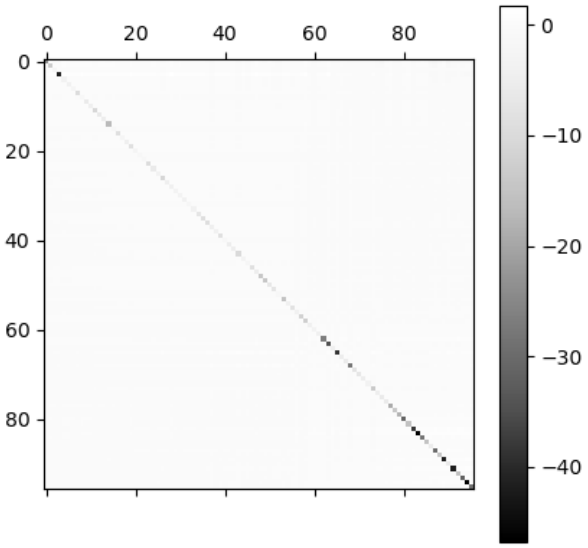

Figure 1.7: Matrix of estimated own and cross price elasticities for the year 1990. The products included are the ones with an estimated own price elasticity above the 25th percentile.

1.12 Appendix: Simulation of the Bertrand Network game

In this section, I perform a simple simulation exercise to summarize and visualize how the Bertrand Network model works.

Table 1.5 describes the parameters used in the simulation. There is a single market with $J = 30$ products and $K = 7$ characteristics whose values are drawn from a uniform distribution in between $[0, 1]$. The demand intercept α is the same across all products and set to 0.15 whereas marginal costs are heterogeneous across products and drawn from a $[.01, .03]$ uniform distribution.

Parameter	Value
J	30
K	7
α	0.15
c	U[0.01, 0.03]
x_{jk}	U[0, 1]

Table 1.5: Parameters for simulation of Bertrand network game

Given this parameters, Figure 1.8 plots the underlying Bertrand network. Each product is a node and the edges capture the degree of substitution between any two products/nodes; the longer the edge the less substitute are the two products. The location of dots and edges is exogenous and entirely determined by the realization of the draws of product characteristics. Conversely, the size of the dots is endogenous and it is proportional to the equilibrium price-cost margins. The plot shows that nodes that are more peripheral tend to have larger dot sizes whereas dots that are more central are smaller. The intuition for this result is the following: peripheral products are more unique or equivalently less central and, per equation (1.13) will charge higher margins in equilibrium. On the other hand, more central nodes face more intense competition and must lower their margins.

Figure 1.9 instead is a visualization of Proposition 3 and plots the equilibrium price-

cost margins on the y-axis against the Bonacich product centrality on the x-axis. It should be clear by now why the relationship is decreasing; higher centrality implies lower equilibrium markups. The noise around the downward sloping relationship is due to the fact that marginal costs are heterogeneous. By increasing the variance of the distribution of costs, Figure 1.9 would start looking noisier and the resulting relationship between centrality and margins might not look as clear. This highlights how empirically it is important to control for the unobserved costs in order to recover the downward sloping relationship. The same would be true if we were to introduce heterogeneity in the demand intercept α.

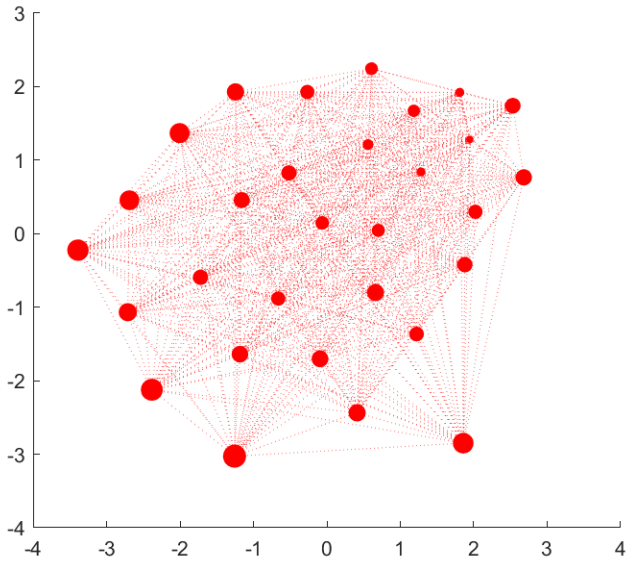

Figure 1.8: Simulated Network. Location is exogenous. Node size is proportional to markups.

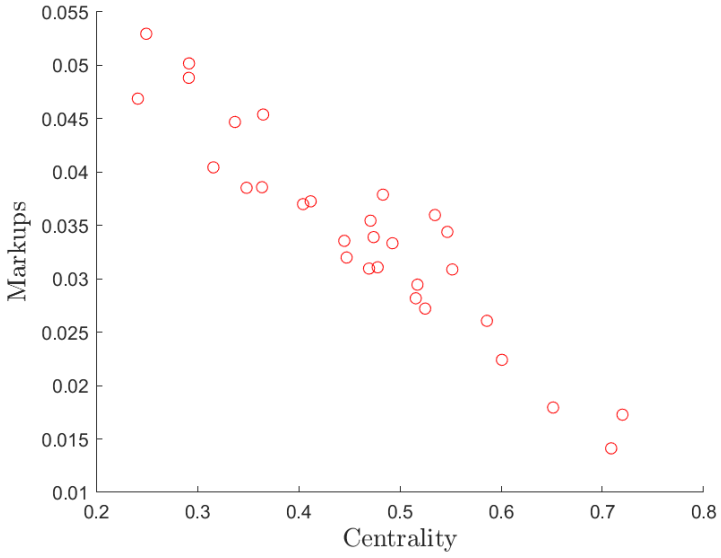

Figure 1.9: Simulated Network. Price-cost margins (y-axis) against Bonacich centrality (x-axis).

1.13 Appendix: From linear-quadratic to discrete choice

This section shows how the linear-quadratic demand model nests the discrete choice demand framework. To see this suppose consumers can only purchase one unit of a product. Under this assumption consumer i problem can be formalized as

$$\max_{(q_{ij})_{j=1}^{J}} \quad q_i'(\alpha_i - p) - \frac{\beta_i}{2} q_i' \left(I_J + \eta X X'\right) q_i \tag{1.93}$$

$$\text{s.t.} \quad q_{ij} \in \{0, 1\} \tag{1.94}$$

$$\sum_{j=1}^{J} q_{ij} \leq 1. \tag{1.95}$$

From the constraints it is immediate to see that the problem boils down to simply choosing the product that provides the highest utility

$$\max_{j \in \{1,\ldots,J\}} \quad \alpha_{ij} - p_j - \frac{\beta_i}{2}(1 + \eta x_j' x_j). \tag{1.96}$$

Next, recalling that $\alpha_{ij} = \eta x_j' \alpha_i^x + \alpha_{ij}^q$, redefining the vectors of product characteristics and preference parameters

$$\hat{x}_j \equiv (x_j', (1 + \eta x_j' x_j)/2)' \tag{1.97}$$

$$\hat{\beta}_i \equiv (\eta \alpha_i, \beta_i)', \tag{1.98}$$

and, rewriting $\alpha_{ij}^q = \hat{\xi}_j + \varepsilon_{ij}$ we can write (1.96) following the standard random utility notation used in empirical applications

$$\max_{j \in \{1,\ldots,J\}} u_{ij} \equiv -p_j + \hat{x}_j' \beta_i + \hat{\xi}_j + \varepsilon_{ij} \tag{1.99}$$

where $\hat{x}_j$ is the vector of product j's characteristics observed by the econometrician, $\hat{\xi}_j$ is

a scalar that captures characteristics or demand shocks unobserved by the econometrician, β_i is a vector of preference parameters capturing how consumer i values different product attributes and, lastly, ε_{ij} is a random utility shock.[36]

By making specific assumptions on the distribution of ε_{ij} one can recover all standard models of discrete choice demand, such as logit, probit, nested-logit and random-coefficient logit.

36. I am implicitly assuming that the econometrician knows or can calibrate η. If this is not the case then the quadratic term should be treated as unobserved and included in the $\hat{\xi}_j$.

CHAPTER 2

PLAN MENUS, RETIREMENT PORTFOLIOS, AND INVESTORS' WELFARE

2.1 Introduction

Employer-sponsored defined contribution (DC) retirement plans are a crucial component of the US savings system, holding nearly $11 trillion in assets as of 2021.[1] These plans allow employees to allocate a portion of their pre-tax income towards retirement savings through a range of investment options, typically mutual funds, to build long-term wealth. For many workers, the assets held in DC plans are among the most important components of their balance sheets and are a significant determinant of their future retirement security.[2]

Despite their importance, many plans do not provide their investors (i.e., employees) with cost-efficient investment options. For instance, in 2019, over half of the plans failed to offer low-cost S&P 500 index trackers. Perhaps even more strikingly, one out of every five plans did not offer an equity fund with a fee below 10 basis points (Figure 2.1).[3]

A closer look at plan expenses reveals substantial dispersion across sponsors, with the difference in the average expense between plans at the 75th and 25th percentiles of about 40 basis points (Figure 2.2). To put this in perspective: assuming an annual return of 6%, if an employee with an annual income of $70,000 contributes 10% to their 401(k) and shifts from

1. Among those, 401(k) are the largest totalling $7.7 trillions. As a share of the US retirement market assets, employer-sponsored DC plans account for 30%. If one includes individual retirement accounts (IRA), DC plans account for 63% of the US retirement assets. https://www.icifactbook.org/.

2. According to the 2019 Survey of Consumer Finances (SCF), for a working age household, the average account balance in a DC plan (including IRAs) was nearly $270,000. More generally, retirement accounts are the second-most commonly held type of financial asset after transaction accounts (https://www.federalreserve.gov/publications/files/scf20.pdf). See also Badarinza, Campbell and Ramadorai (2016).

3. On average, around 50% of plans do not offer the cheapest share class of a fund even though they meet its minimum investment requirement (Figure 2.10).

a plan at the 75th percentile to one at the 25th percentile, they could save approximately $95,000 in investment fees.[4]

Beyond dispersion, plan expenses are also surprisingly high, with the asset-weighted average expense ratio for the median plan in the 2019 cross-section close to 40 basis points.[5] For context, in that same year, had a retail investor constructed a portfolio of Vanguard index funds to obtain exposure to all asset classes available in a typical retirement plan, the expense ratio would have been more than four times lower.[6] Additionally, more expensive plans do not produce better investment performance for their investors (Figure 2.3).[7] All things considered, it is unsurprising that employees have increasingly sought to hold plan sponsors (i.e., employers) accountable for allegedly violating their fiduciary duties, with high investment costs emerging as the common theme in many recent 401(k) lawsuits (Mellman and Sanzenbacher (2018)).[8]

Why do many plans feature investment options that are less cost-efficient than comparable alternatives in the marketplace? And can policy regulating the design of retirement plans effectively help reduce these costs and improve investors' outcomes?

To address the first question we need to understand what drives sponsors' decisions to include high-cost funds in their retirement plan. Although sponsors have a fiduciary duty

4. The calculation assumes a working period of 40 years and that the annual return is the same. In practice, plans with higher expenses tend to have gross of fees returns that are even lower (Table 2.5).

5. These patterns are not limited to the 2019 cross-section. Plan expenses were even higher before 2019. At the same time, the dispersion in plan expenses has been roughly stable over time (Figures 2.11) and remains even when comparing plans of similar size (Figures 2.12).

6. In 2019, the expense ratio for a Vanguard equally-weighted portfolio of retail index funds, including its International equity index funds (VEMAX, VEUSX, VPADX), US Equity funds (VGSLX, VFIAX, VIMAX, VSMAX) and Bond Fund (VBTLX) is below 10 basis points. By retail, I mean that the minimum investment required is none or limited. Figure 2.13 compares the expense for the median plan against this portfolio of Vanguard index funds over time.

7. Table 2.5 and Figure 2.3 show that the plan-level (gross of fees) performance tends to be lower for more expensive plans. Similar patterns have been found in the context of active investing (Gil-Bazo and Ruiz-Verdu (2009)) contrasting what frictionless models with rational investors predict (Berk and Green (2004)).

8. I provide more details of some recent 401(k) lawsuits in Appendix 2.14.

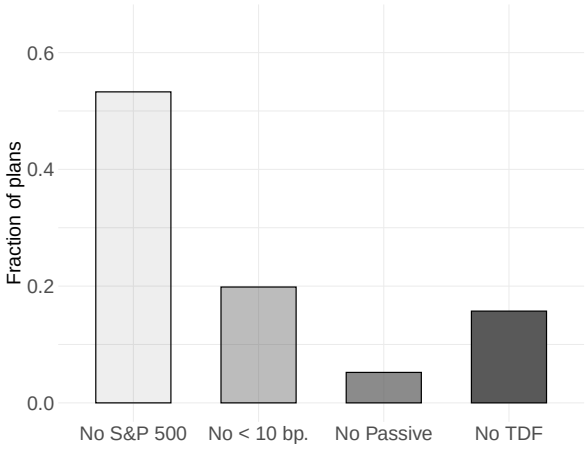

Figure 2.1: Plan quality for the year 2019. 'No S&P 500' is the share of plans without an S&P 500 tracker with a fee below 10bp.

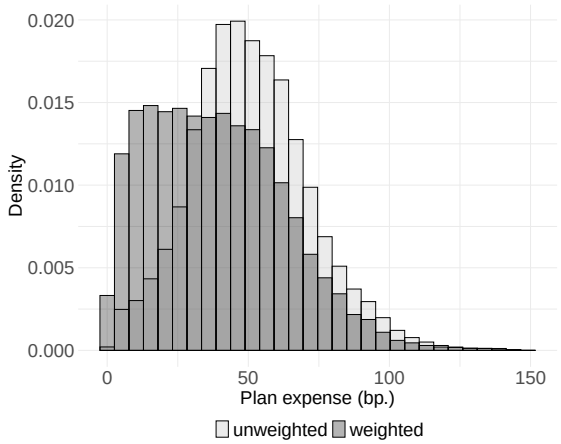

Figure 2.2: Distribution of asset-weighted and unweighted average plan expense ratio across plans for the year 2019.

to design their plan in the investors' best interest, agency frictions may reduce sponsors' sensitivity to funds' fees and make them value attributes other than fees when constructing their plan menu. For example, sponsors may favor the inclusion of funds affiliated with their plan provider (a.k.a recordkeeper, Pool, Sialm and Stefanescu (2016)) or may favor the inclusion of costlier funds to reduce direct fees paid to the recordkeeper (Badoer, Costello and James (2020), Bhattacharya and Illannes (2022), Pool, Sialm and Stefanescu (2022)).[9]

Addressing the second question requires understanding how investors allocate their contributions across the options available in their plan. Their investment behavior is crucial to evaluate policies regulating the design of retirement plans. For instance, if many investors are inactive because of inertia (Madrian and Shea (2001), Beshears, Choi, Laibson and Madrian (2009), Choi (2015)), policies mandating the inclusion of low-cost funds, like an S&P 500

9. Plan sponsors outsource administrative tasks such as maintaining employees' account balances to plan providers (a.k.a recordkeeper), which are often vertically integrated into fund provision. Employers can compensate plan providers either directly or indirectly by offering expensive funds that pay higher kickbacks. Part of this compensation can also cover for services, like financial advising, that I do not observe. Differences in unobserved services can explain the observed dispersion in plan expenses. Figure 2.14 shows that the dispersion in plan expenses remains substantial even when comparing employers within the same industry, size and plan provider. This suggests that variation in unobserved services across providers or employers cannot be the major source of differences in plan expenses.

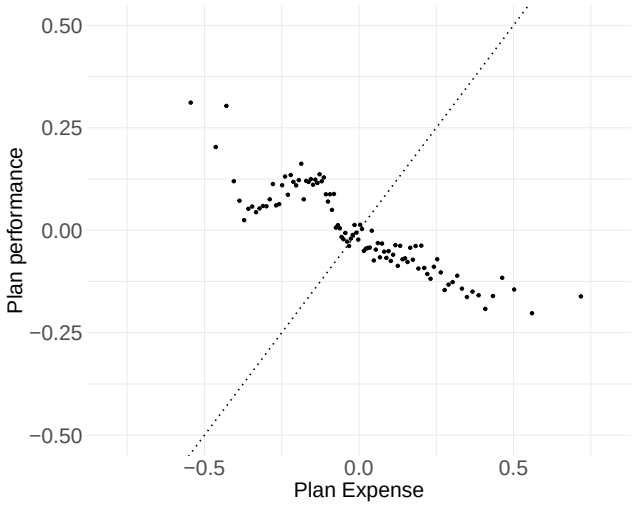

Figure 2.3: Binned scatter of plan (gross) performance and average plan expense corresponding to the specification in the second column of Table 2.5. Expenses and performance are yearly demeaned and measured in percentage points.

tracker, might be ineffective. This is because inactive investors are unlikely to reallocate their contributions to the new low-cost option.

From a supply-side perspective, funds' competition incentives are key to rationalize excessive fees and to predict how these fees will respond to policy changes. Understanding competition in this space requires a model that accounts for how funds set fees in response to sponsors' incentives to include them in their plan and in response to investors' incentives to substitute toward competitors' funds available in the plan.

In this paper, I combine tools from finance and industrial organization to develop a structural model of plan menu choice, retirement portfolio choice and fee competition between funds providers. The model features a two-layer demand system where, in the first layer, sponsors construct their retirement plan and, in the second layer, plan investors form their retirement portfolio from the options available in their plan menu.

Plan sponsors compose their menu by selecting investment options from the pool available in their recordkeeper's network of funds.[10] In modelling sponsors' preferences, I accommo-

10. When forming their plan menus sponsors might not be aware of all available options. Part of this is

date for two types of agency frictions. First, sponsors have a taste for funds' affiliation. Everything else equal, a fund affiliated with the plan recordkeeper will be more likely to be included in plan menus. Second, sponsors' fee sensitivity can be arbitrarily misaligned to that of investors. This captures how the compensation structure between sponsors and recordkeepers might incentivize the former to favor the inclusion of expensive funds. I model sponsors' menu choice as a two-stage multiple discrete choice problem. First, sponsors decide whether to include or not a particular investment category. Once this decision is made, they proceed to choose the options within that category that offer the highest indirect utilities. Sponsors face some cost to adding more than one option within each category, which the model captures by assuming that the number of options included is drawn randomly with decaying probability.[11]

In the second layer, plan investors allocate their contributions among the options available in their menu. Unlike in much of the prior IO literature, investors preferences are tied to a standard portfolio problem, with risk aversion over the variance of returns. These preferences drive substitution patterns between funds, and unlike in most structural IO demand models, this framework generally allows funds to be either complements or substitutes to each other. In modelling how this type of demand system determines funds' fee-setting incentives, I rely on tools I develop in a related paper (Loseto (2023)), which exploits the close relationship between standard demand systems and the network games literature (Ballester, Calvó-Armenagol and Zenou (2006)). To capture investors' inertia, I allow for the possibility

because different recordkeepers have different pre-existing relationships with specific mutual fund providers. In fact, plan-level data indicates that recordkeepers' networks of funds are far from perfectly overlapping (Figure 2.15).

11. Adding options may be costly for several reasons. First, sponsors fiduciary duty is not limited to the design of the retirement menu but also requires them to monitor the included options and provide plan investors with up-to-date information about their performance. Second, sponsors' asset base (i.e., the total contributions) is limited, and with too many options, it could be challenging to meet the minimum investment requirements required by investment funds. In estimation, I calibrate this probability to match the observed distribution of the number of funds included within each category (Figure 2.9). In Appendix 2.11, I offer a simple microfoundation for the optimal choice of the number of options to be included within each category building on the Stigler (1961) simultaneous search model.

that some investors do not make an active investment decision. Instead, they default their asset allocation into the plan default option, typically a Target-Date fund (TDF).

I estimate sponsors' and investors' preferences using comprehensive plan-level data. This information is collected from 401(k) plan menus and asset allocations as reported by plan sponsors on form F5500 to the Department of Labor (DOL).[12] Specifically, I compute the probability of a given investment fund being included in a plan from the observed plan menus. The observed variation in these inclusion probabilities enables me to identify and estimate plan sponsors' preferences. Similarly, the observed variation in plan-level asset allocations allows me to identify plan investors' preferences. To account for the possible endogeneity of investment fees, I exploit the granularity of the data to control for unobserved demand shocks along several dimensions, including investment category fixed effects, sponsors fixed effects, funds' brand fixed effects and a fixed effect for passive funds. To further ensure that the residual variation in fees is uncorrelated with unobserved demand shocks, I use two types of cost-shifter instruments: funds' turnover ratios and a Hausman-type instrument. Funds' turnover measures trading-related transaction costs passed on to investors through higher fees (Pástor, Stambaugh and Taylor (2020)).[13] As for the Hausman strategy, it uses the average fee charged by funds within the same fund provider but from other investment categories as an instrument for funds' fees (Hausman (1996)).

Model estimates indicate that plan sponsors are less sensitive to fees than plan investors. This is particularly evident when comparing them to investors actively forming their retirement portfolios, who are over twice more elastic to fees than sponsors.[14] This misalignment in sponsors' and investors' elasticity to fees suggests that sponsors may not adequately inter-

12. See Section 2.3 for more details on the data.

13. Pástor, Stambaugh and Taylor (2020) show how funds' profit-maximizing behaviour implies that funds' optimal fee must depend on funds' trading costs which are a function of funds' turnover.

14. In Appendix 2.13 I model sponsors' preferences as a weighted average of their true preferences and investors' preference and estimate that sponsors weigh their own preferences nearly three times more than investors' preferences.

nalize their employees' preferences when constructing retirement plan menus. The estimates also indicate that sponsors have a strong preference for including funds affiliated with their recordkeeper. Quantitatively, being affiliated with the plan recordkeeper increases the inclusion probability by 0.36 percentage points, a magnitude four times higher than what a 10 basis points reduction in funds' fees would lead to.

I model supply as a differentiated Bertrand oligopoly where funds providers set fees simultaneously before sponsors make their plan menu decisions and plan investors form their portfolios. Funds compete along two margins. First, they compete to be included in sponsors' plan menus and internalize how agency frictions reduce sponsors' fee elasticity. Second, conditional on inclusion, they compete for plan investors' assets. When setting fees, they internalize investors' inertia, and that the likelihood of facing close competitors is low because sponsors tend to include no more than one fund in each investment category (Figure 2.9).[15]

After estimating sponsors' and investors' demands, I use the Nash-Bertrand equilibrium conditions to recover funds' marginal costs and price-cost margins. The median fund, spanning all fund types, charges a margin of 14 basis points. Given the median expense ratio, this translates to a markup of approximately 20%. Looking at passive funds and TDFs specifically, their estimated marginal costs indicate that these funds are more efficient than others. However, they do not fully transfer these efficiency gains to investors. This is especially evident for TDFs, where the estimated median markup stands at about 39% nearly twice the median markup observed across all funds.

TDFs' pricing power arises from two forces. First, they are often funds affiliated with

15. While most sponsors aim to provide investors with a broad range of investment categories (e.g., Large Cap Value, etc.) they typically offer only one fund per category (Figure 2.9). Although this plan structure helps manage risk, it could weaken competition between funds, as price competition is more intense when products are more alike e.g., when funds belong to the same category. Consistent with this Bertrand-type of intuition, plan-level data suggests that fees are inversely correlated with the number of options offered within a specific category (Table 2.6).

the plan recordkeeper. Sponsors value this affiliation, making these funds more likely to be included in plan menus. Second, following the 2006 Pension Protection Act, they were designated as a qualified default option (QDIA) for employer-sponsored retirement plans, allowing them to attract demand from inactive investors who are inelastic to fees. A recent study by Vanguard finds that the share of plan investors holding a single TDF increased from 20% in 2010 to 54% in 2019, suggesting that the fraction of inactive investors might be far from negligible.[16] Model estimates match this evidence closely, indicating that two out of five investors did not make an active investment decision and that the fraction of inactive investors more than doubled over time from around 25% in 2010 to nearly 60% in 2019.

In the last part of the paper, with the estimated demand and supply parameters, I explore a series of counterfactuals to quantify the effects of plan design policies on plan investors' welfare. In the first counterfactual, I eliminate agency frictions whereby plan sponsors favor the inclusion of funds affiliated with their recordkeeper. This policy turns out to be ineffective: sponsors simply substitute from affiliated funds to similarly expensive unaffiliated funds and overall costs for investors do not meaningfully decrease. In other words, eliminating sponsors' preference for affiliated funds does not make them more responsive to fees.

In the second set of counterfactuals, I consider policies that mandate the inclusion of a low-cost equity index fund tracking the S&P 500 and the inclusion of a low-cost TDF. Mandating the inclusion of low-cost equity S&P 500 funds leads to an increase in investors' surplus of about 2%. The increase is modest because investors value lower fees, but they also want to diversify across the available funds, dampening the incentive to substitute toward the low-cost index fund. Moreover, this policy only benefits active investors, leaving inactive investors' surplus unchanged. Mandating the inclusion of low-cost TDFs increases investors' surplus by about 11%, a magnitude more than five times larger than the previous policy.

16. The study is based on an examination of retirement plan data from 5 million defined contribution plan participants across Vanguards recordkeeping business. See *'How America Saves'*, Vanguard (2022).

This policy is more effective because it also benefits inactive investors who reallocate their entire portfolio to the low-cost TDF.

Lastly, I consider a policy that caps funds' expense ratios at 50 basis points. Under this policy, sponsors can only offer funds with expenses below this cap.[17] Investors' outcomes further improve, with an increase in surplus of approximately 14%. This policy is effective because it affects the entire menu of options by limiting the inclusion of the most inefficient funds. It benefits inactive investors because it replaces the most inefficient TDFs, and it benefits active investors by reducing the costs of all options available in the plan menu.

The rest of the paper proceeds as follows. Section 2.2 describes how this paper contributes to the literature. Section 2.3 describes the data. Section 2.4 sets up the demand side of the model. Section 2.5 focuses on the supply side and characterizes the equilibrium fees. Section 2.6 estimates the demand side of model. Section 2.7 turns to the supply side and recovers funds' price-cost margins. Section 2.8 presents the results of policy counterfactuals and Section 2.9 concludes.

2.2 Contributions to the Literature

This paper contributes to the literature that studies retirement investing and the design of retirement plans. A large part of this literature has focused on the demand side by studying 401(k) enrollment decisions with a particular focus on the role of automatic enrollment into default options (Madrian and Shea (2001), Beshears, Choi, Laibson and Madrian (2009), Carroll, Choi, Laibson, Madrian and Metrick (2009), Choi (2015)), behavioural biases in retirement investing (Benartzi and Thaler (2001), Huberman and Jiang (2006), Benartzi and Thaler (2007), Tang, Mitchell, Mottola and Utkus (2010)), the demand for financial advice (Chalmers and Reuter (2020), Reuter and Richardson (2022)), and the implications

17. Under this policy, most of active funds will not find profitable to operate in the retirement market. Nevertheless, investors' surplus improves because most of these funds do not generate enough alpha to justify their expenses (Jensen (1968), Gruber (1996), Carhart (1997), Fama and Kenneth (2010)).

of automatic enrollment on saving behaviour over the life-cycle (Choukhmane (2021), Duarte, Fonseca, Goodman and Parker (2022)).

Another part of the literature has looked at the supply side and has mainly focused on empirically examining the role of agency frictions between plan providers (i.e., recordkeepers) and plan sponsors (i.e., employers). Pool, Sialm and Stefanescu (2016) show that plan providers vertically integrated into fund provision tend to favor affiliated funds, Badoer, Costello and James (2020) provide evidence of how plan providers trade-off direct fees from the sponsor with indirect fees paid by funds via revenue-sharing agreements and Pool, Sialm and Stefanescu (2022) show that funds paying revenue-sharing fees are more likely to be included in a plan and less likely to be deleted. More recently, Gropper (2023) shows how employers distort plan menus to reduce litigation risk. They do so by reducing the number of options offered, which reduces employees' retirement wealth. Building on this evidence, Bhattacharya and Illannes (2022) take a more structural perspective and develop a model where revenue-sharing fees are the outcome of a bargaining game between sponsors and recordkeepers. Yang (2023) instead models employers' dynamic decision to switch plan provider and shows how switching costs can rationalize a sizable share of the observed dispersion in plan expenses.

This paper lies in between these two strands of work. Its primary contribution is to develop and estimate a structural model of plan menu choice, retirement portfolio choice and fee competition between differentiated fund providers. I model sponsors' menu choice as a multiple discrete choice problem building on the workhorse discrete choice frameworks developed in Berry (1994) and Berry, Levinsohn and Pakes (1995). I accommodate for agency frictions by allowing sponsors' preferences to depend on funds' characteristics that relate to the identity of the plan recordkeeper, without modelling the latter as separate agents. For example, I allow sponsors to have a taste for whether or not a fund is affiliated with its recordkeeper. Moreover, by modelling and identifying sponsors' and investors' preferences

separately, I can allow their sensitivity to expenses to be arbitrarily misaligned and capture agency frictions whereby sponsors favor expensive funds to reduce the direct fees paid to the recordkeeper.

A key goal of the paper is to quantify the effects of alternative plan design policies on investors' welfare. To this end, I incorporate investors' portfolio decisions into the model. After sponsors' menus have been chosen, I assume plan investors with quadratic preferences (Markowitz (1952)) form their retirement portfolio from the options available in their menu. In a recent contribution, Egan, MacKay and Yang (2023) also model retirement portfolio choice as a mean-variance problem and show how variation in expense ratios can identify investors' beliefs and risk aversion separately. I complement their work in two ways. First, I extend their methodology by also allowing investors to be inactive with some probability, in which case, investors' contributions are defaulted into some of the available TDFs. Second, I develop the supply side and characterize the Nash-Bertrand equilibrium fees when investors have mean-variance demand rather than discrete choice demand.

I assume funds compete in a differentiated oligopoly by setting fees simultaneously a la Nash-Bertrand and use the implied first-order conditions to recover funds' marginal costs and price cost margins. This links my paper to the empirical industrial organization literature on imperfect competition in the mutual fund industry. Most of this literature models investment decisions on the demand side as a standard discrete choice problem. This is done either because the focus is on competition between financially homogeneous products such as S&P index funds (Hortaçsu and Syverson (2004)), ESG and non-ESG funds that track the same underlying index (Baker, Egan and Sarkar (2022)) and variable annuities (Koijen and Yogo (2022)), or because investors are assumed to be risk neutral (Massa (2003), Roussanov, Ruan and Wei (2021)).

No paper in this literature has modeled fee competition between differentiated funds when investors have mean-variance preferences and characterized the resulting equilibrium

fees. To the best of my knowledge this is the first paper to do so. Specifically, I show how the quadratic structure of investor preferences implies that equilibrium fees can be decomposed into three components. One of these components, which I refer to as 'Hotelling markdown', is equivalent to the vector of Bonacich centralities of a network in which funds are the network nodes and funds' substitution patterns are (inversely) proportional to the network edges. This markdown captures how much a monopolist should give up when it faces competitors that are closer in the space of product characteristics. A more central fund faces more similar competitors and must charge lower fees (e.g., has a higher Hotelling markdown).

In a related paper (Loseto (2023)), I develop part of this decomposition in the context of price competition between multi-product firms selling differentiated products to consumers with quadratic preferences. Building on a growing literature on linear-quadratic network games (Ballester, Calvó-Armenagol and Zenou (2006), Ushchev and Zenou (2018), Pellegrino (2023)), I show how a firm's Bonacich network centrality fully summarizes its pricing power: firms that are more central charge lower prices because their products are not sufficiently differentiated from their competitors products. The characterization of the equilibrium fee I provide in this paper is more general because the network is ex-ante unknown to players (e.g., funds do not know which plan will include them), and players' actions influence the resulting network structure (e.g., when setting fees, funds influence their plan inclusion probability).[18]

Lastly, my paper joins a growing literature at the intersection between industrial organization and financial economics that uses structural models to study market structure and competition. This literature has looked at mortgages (Allen, Clark and Houde (2014), Benetton (2021), Robles-Garcia (2021), Agarwal, Grigsby, Hortaçsu, Matvos, Seru and Yao

18. Grice and Gueciouer (2023) also study fee competition between investment providers from a network perspective. Building on Grice (2023), they propose a model of competition between fund families where the competitive network is micro-founded from investors' imperfect consideration. My model instead features competition at the fund level and belongs to the class of network games studied in Loseto (2023) where the underlying network is determined by assets' characteristics, thereby allowing for arbitrary substitution patterns between products. Moreover, in the model I am considering, the network structure is unknown to players because investment providers set fees before plan sponsors design their retirement menu.

(2021)), credit cards (Nelson (2020)), loans (Cuesta and Sepúlveda (2020), Benetton, Buchak and Robles-Garcia (2022)), banking (Egan, Hortaçsu and Matvos (2017), Buchak, Matvos, Piskorski and Seru (2018), Buchack, Matvos, Piskorski and Seru (2022)), municipal bonds (Brancaccio, Li and Schüroff (2020)), auctions (Hortaçsu and McAdams (2010), Kastl (2011), Richert (2021)), variable annuities (Koijen and Yogo (2022)) and the market of financial advice (Egan (2019), Bhattacharya, Illanes and Padi (2020)).

2.3 Institutional Setting and Data Sources

In this section I describe what is an employer-sponsored retirement plan and overview its administrative structure. After that I describe the data sources and provide some summary statistics about my sample of retirement plan menus and the investment options available therein.

2.3.1 What is an employer-sponsored retirement plan?

Employer-sponsored retirement plans are savings vehicles designed to assist employees in accumulating wealth for their retirement years. Although there are various types of employer-sponsored retirement plans, the most common is the defined contribution plan. At its core, a defined contribution plan is a retirement plan in which an employee makes regular contributions. The final amount available upon retirement is not pre-determined but instead depends on the contributions made and on the returns obtained from the investment options available in the plan.

Under a DC plan, employees contributions represent a percentage of their salary which is subtracted from their paycheck before taxes, thereby reducing their current taxable income. The gains in the retirement account grow tax-deferred, implying taxes are not due until funds are withdrawn during retirement years. Withdrawals from a DC plan before a certain age (typically between 59 and 60) can result in penalties. After reaching retirement

age, participants might withdraw distributions as lump sums, period payments or annuities. Importantly, all withdrawals at this point are subject to standard income tax.

Employers play a crucial role in the provision and design of these type of plans. First, many employers offer to match a portion of the employee's contributions. For example, an employer might contribute 50 cents for every dollar the employee contributes, up to a certain percentage of the employee's salary (typically around 6%). More importantly, employers are in charge of selecting and monitoring which investment options, typically mutual funds, are to be included in the plan. Once the employee decides what percentage of their income to contribute, they typically have the autonomy to allocate their contributions, together with the part matched by their employer, across the options available within their plan. In many cases, to help less financially savvy employees or those who do not make an active investment choice, plan sponsors include options, such as TDFs, known as Qualified Default Investment Alternative (QDIA) to be used as default option when an employee contributes to the plan without specifying how their contribution should be invested.

Figure 2.4 summarizes graphically the structure of a DC plan. Plan sponsors typically hire recordkeepers to assist in the design of their retirement menu and to perform administrative tasks such as maintaining account balances. Most recordkeepers, like for example Fidelity, are also providers of investment funds and it is not uncommon for retirement plans to include funds from the recordkeeper product line. Overall, a retirement plan consists in a set of assets, spanning a broad range of investment categories, selected by sponsors and recordkeepers. Under the Employee Retirement Income Security Act (ERISA), plan sponsors are fiduciaries to their employees and are subject to litigation risk if their retirement plan is not designed in the employees' best interest. The inclusion of high-cost investment options or the lack of low-cost options are among the main triggers of some recent lawsuits, as I describe in Appendix 2.14.

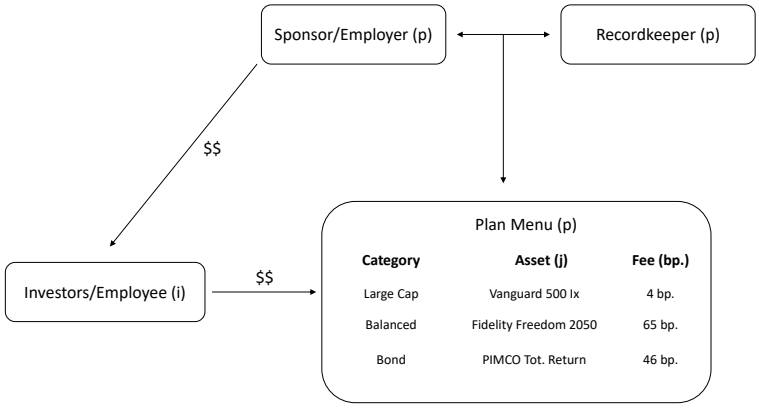

Figure 2.4: Administrative structure of a defined-contribution employer-sponsored retirement plan.

2.3.2 Data

The primary data source for this study is collected from Form 5500. This form is annually filed by employers with the Department of Labor (DOL) in adherence with the ERISA regulations. Within this form, Schedule H provides detailed information about retirement plan menus. Specifically, it contains data regarding the investment options offered within an employer's retirement plan and the plan-level dollar allocation to each of these options.

Although Form 5500 filings are available for download from the DOL website, information about plan menus comes in a non-standardized format, stored in pdf images, that would need to be digitized manually. I acquired the digitized version of these filings for the years 2010 to 2019 from BrightScope Beacon who collects and digitizes these data directly from the publicly available DOL filings. Overall, the data covers more than 90% of total plan assets in each year from 2010 to 2019, digitizing an average of 55,000 plans per-year.[19]

I complement this data with additional information from the DOL Form 5500, including the number of plan participants and the identity of each plan recordkeeper. Also, I obtain

19. See Table 2.7 for more details on the data coverage.

	N	Mean	SD	p5	p25	p50	p75	p95
Tot. Assets (mln.)	60798	27.85	59.15	1.78	5.55	11.15	25.11	107.64
N. of participants	60798	475.19	597.08	118.00	160.85	250.10	494.83	1723.87
N. of options	60798	24.72	14.46	12.96	17.88	21.55	26.91	47.86
N. of categories	60798	14.96	3.27	9.55	13.00	15.00	17.01	20.00
Affiliated (%)	60798	25.13	27.56	0.00	0.64	14.33	43.24	82.33
Target Date (%)	60798	78.36	37.63	0.00	73.17	100.00	100.00	100.00
Passive (%)	60798	92.17	23.67	17.35	100.00	100.00	100.00	100.00
Avg. expense (pp.)	60798	0.63	0.21	0.26	0.49	0.63	0.76	0.98
Avg. expense (w.)	60798	0.51	0.24	0.11	0.35	0.51	0.66	0.93
Assets per participants ($ thous.)	60742	50.50	60.80	4.84	17.50	35.05	64.67	143.73
Equity	60798	0.60	0.09	0.46	0.56	0.61	0.66	0.73
Bond	60798	0.18	0.06	0.09	0.14	0.18	0.21	0.28
Balanced	60798	0.17	0.09	0.02	0.13	0.17	0.22	0.31
N. of years	60798	5.33	2.90	1.00	3.00	5.00	8.00	10.00
N. of recordkeepers	60798	1.20	0.47	1.00	1.00	1.00	1.00	2.00

Table 2.1: Plan level summary statistics for the years 2010 to 2019. Each variable is first averaged within plan across years and then tabulated across plans. The variable 'N' is the number of plans and the variable 'N. of years' is the number of years a plan is observed in the sample. Sample is for sponsors with number of participants between 100 and 5000.

data on funds' historical expense ratios from the Center for Research in Security Prices (CRSP) and merge those in the main dataset using funds' tickers. BrightScope also provides data on funds' expense ratios, but the historical expenses are only available starting from 2016.[20]

Table 2.1 provides some summary statistics for the plans in my sample. Following Bhattacharya and Illannes (2022), I focus on plans whose number of participants is between 100 and 5000, representing roughly 95% of the whole sample.[21] The average plan has close to 30 millions dollars in assets and around 475 participants. Both measures of plan size are right-skewed due to the presence of large plans, with the median plan having assets ranging around 10 millions and 250 plan participants.

20. Before 2016, BrightScope reports the most recent funds' expense ratio which I replace with the one obtained from CRSP.

21. Vary large plan sponsors are more likely to engage in private negotiations with recordkeepers and mutual fund providers to obtain fees that are lower than fees available in the mutual fund market. These privately negotiated fees are not available in the BrightScope data.

Looking at the characteristics of the investment menu, the average plan offers 25 investment options across 15 different investment categories. For the average plan, one out of four options is affiliated with the plan recordkeeper. Moreover, most of the plans offer at least a TDF fund and a passive fund. Turning to plan expenses, the average expense ratio charged by the average plan is of about 63 basis points. This is more than 10 basis point higher than its asset-weighted average, suggesting that plan investors tilt their allocations toward cheaper funds.

Table 2.8 reports some summary statistics for the funds offered in the sample of plan menus I observe. Excluding cash accounts and common stocks, there are 5600 distinct funds for which at least a ticker identifier is available.[22] In my sample of plans, the average fund manages 191 million dollars of retirement assets and has an average portfolio share of about 3%. The distribution of assets is right-skewed, with the median fund having 7 millions in assets. Interestingly, the share of plan assets allocated into a given fund varies significantly across plans, with the average standard deviation of the portfolio share across funds being as large as 3%.

The data also suggests that sponsors review their menu of investment options often. In Table 2.8, the penultimate row reports the average fraction of years a fund is included within a plan menu. On average, a fund is offered in only half of the years that I observe the plan menu, implying that sponsors regularly modify their menu offerings. The same does not appear to be true when looking at plans' recordkeepers, with more than 75% of plans having the same recordkeeper over the sample period (Table 2.1).

2.4 Demand

This section describes the two-layer demand side of the model. I start from the first layer where I describe sponsors' preferences and the menu choice problem they face. After that, I

22. The same fund may have multiple tickers, one for each different share class.

derive funds' plan inclusion probabilities implied by sponsors' demand. Next, I turn to the second layer of demand where I first describe investors' preferences and then derive individual and aggregate asset demand systems.

2.4.1 Sponsors' menu choice problem

Throughout the paper I will index plan sponsors by p and mutual funds by j. All vectors are in bold. The goal of sponsor p is to choose a set of mutual funds to include into its retirement plan. Typically sponsors hire recordkeepers to help in designing and managing their plan menu and, empirical evidence suggests that the set of funds sponsors consider to being with is strongly influenced by the recordkeeper identity. From a modelling point of view, I capture this by allowing for heterogeneity in sponsors' consideration sets.[23] Formally, I assume that sponsor p considers with positive probability a subset $N_p \subset N$ of all mutual funds available. Empirically, I assume that fund j belongs to N_p if I observe at least one plan that includes j and shares the same recordkeeper as p. In other words, sponsors with the same recordkeeper have the same consideration set.

Sponsor p's random utility from including fund j is given by

$$u_{jp} = V_{jp}(\boldsymbol{\theta}_p) + \varepsilon_{jp} \tag{2.1}$$

with the non-random utility part, V_{jp}, defined as

$$V_{jp}(\boldsymbol{\theta}_p) = \boldsymbol{w}'_{jp}\boldsymbol{\theta}_p + \zeta_j \tag{2.2}$$

where $\boldsymbol{w}_j$ is the vector of fund j's observed characteristics including its expense ratio, past returns and an indicator for whether j is a fund affiliated with p's recordkeeper, ζ_j captures

23. In principle sponsors can choose/change their recordkeeper. Yet, the data suggests that sponsors tend to stick with the same recordkeeper over time (Table 2.1). This motivates why I to abstract from modelling recordkeepers and sponsors as separate agents.

characteristics, possibly correlated with fees, unobserved to the econometrician and, ε_{jp} is a random preference shock distributed as T1EV. When modeling the preferences of sponsors I place no restrictions on their parameters, $\boldsymbol{\theta}_p$, thereby allowing for such preferences to be arbitrarily misaligned from those of plan investors. In Appendix 2.13 I show how plan investors' preferences can be nested into sponsors' preferences and how to interpret the parameters $\boldsymbol{\theta_p}$ as a weighted average between sponsors' true preference parameters and investors' preference parameters.[24]

The preference specification in (2.1) captures two types of agency frictions that have been documented in the literature. First, the indicator for funds' affiliation allows for the possibility that sponsors prefer to include funds affiliated with the plan recordkeeper (Pool, Sialm and Stefanescu (2016)).[25] If the preference coefficient on the affiliation indicator is positive, an affiliated fund will be more likely to be included than an otherwise identical unaffiliated fund. Second, sponsors might be willing to include expensive funds to reduce their direct payments to the recordkeeper (Badoer, Costello and James (2020), Bhattacharya, Illanes and Padi (2020)). If this incentive is strong enough, it will affect the estimated preference coefficient on funds' fees and will likely reduce sponsors' elasticity to funds' fees.

I assume that funds are classified into investment categories indexed by $g \in \{1, ..., G\}$ and model the menu choice as a two-stage decision problem. In the first stage, sponsors choose which category to offer in their plan and this decision is made independently across categories. For each selected investment category, in the second stage, sponsors evaluate the options available making within category comparisons and selecting the options providing the highest indirect utility.

To complete the decision problem, I need to specify how the number of options to be

24. Identifying and estimating how much sponsors weigh their investors preferences requires additional assumptions. In Appendix 2.13 I provide more details and find that sponsors weigh their own preferences three times more than their investors preferences.

25. The data suggests that affiliated fund are nearly twice more likely to be included in a plan, even when comparing employers within the same industry and with similar size (Figure 2.19).

included within each selected category is chosen. I assume that this number is drawn randomly from a geometric distribution with parameter q, rather than modelling this choice as the outcome of a rational decision problem.[26] This implies that the probability of n options being included in a given category is given by:

$$q(n) \equiv q(1-q)^{n-1} \quad \text{for } n = 1, 2, ... \tag{2.3}$$

This modelling assumption is guided by the empirical observation that sponsors tend not to include more than one option per investment category. Figure 2.9 plots the empirical distribution of the number of options offered within investment category and shows how in nearly 70% of instances sponsors only include one option per category.[27] The probability then decays as the number of options included increases, consistent with the presence of some cost that sponsors incur when adding more than one option within the same category. In estimation, I calibrate q to match the observed empirical distribution allowing for heterogeneity at the year-recordkeeper-category level.

2.4.2 Funds' plan inclusion probabilities

In this section I derive funds' inclusion probabilities implied by sponsors' menu choice problem described in the previous section. To this end, I will analyze sponsors' decision problem backward.

Consider sponsor p who has chosen to offer category g and needs to select n investment funds within g. At this stage, p ranks all the options according to (2.1) and selects the n

26. As I explain in Appendix 2.11, where I offer a simple microfoundation for the optimal choice of the number of options included within an investment category, incorporating such decision in the full model would considerably complicate its estimation.

27. For each year-plan-category pair I count the number of funds offered and plot the resulting distribution in Figure 2.9.

options $\{j_1, ..., j_n\}$ such that

$$u_{j_1} > u_{j_2} > ... > u_{j_n}. \tag{2.4}$$

Fund j will be included in p's plan if and only if u_j is ranked among first nth highest utilities. Letting j_z be the option with the zth highest utility, the probability that j is included in plan p is given by

$$\phi_{jp}^{1:n} \equiv \sum_{z=1}^{n} \phi_{jp}^{z}$$

where ϕ_{jp}^{z} is the probability that j is ranked in the zth position i.e.,

$$\phi_{jp}^{z} \equiv \Pr\{j = j_z\}.$$

Under the assumption that the random utility shocks are distributed as T1EV, an analytical expression for each ϕ_{jp}^{z} can be derived. For $z = 1$, ϕ_{jp}^{1} corresponds to the standard logit choice probability

$$\phi_{jp}^{1} = \frac{\exp(V_j(\boldsymbol{\theta}_p))}{\sum_{k \in g} \exp(V_k(\boldsymbol{\theta}_p))}. \tag{2.5}$$

For $z = 2$, ϕ_{jp}^{2} is the probability that j provides the 2nd highest utility which equals the sum of probabilities of all utility rankings where u_j is the 2nd largest utility. In a world in which there are only 4 options, say $\{j, k, l, m\}$,

$$\begin{aligned}
\phi_{jp}^{2} = {} & \Pr\{u_{kp} > u_{jp} > u_{lp} > u_{mp}\} + \Pr\{u_{kp} > u_{jp} > u_{mp} > u_{lp}\} \\
& + \Pr\{u_{lp} > u_{jp} > u_{kp} > u_{mp}\} + \Pr\{u_{lp} > u_{jp} > u_{mp} > u_{kp}\} \\
& + \Pr\{u_{mp} > u_{jp} > u_{kp} > u_{lp}\} + \Pr\{u_{mp} > u_{jp} > u_{lp} > u_{kp}\}.
\end{aligned}$$

In Appendix 2.11 I show that the independence of irrelevant alternatives (IIA) property of the logit model implies ϕ^2_{jp} can be written as follows for an arbitrary number of options:

$$\phi^2_{jp} = \sum_{j_1 \neq j} \frac{\exp(V_{j_1 p})}{\sum_k \exp(V_{kp})} \frac{\exp(V_{jp})}{\sum_{k \neq j_1} \exp(V_{kp})}.$$

The above expression can be further generalized that to the case in which we want to compute the probability of j being ranked in an arbitrary position zth

$$\phi^z_{jp} = \sum_{(j_1,\ldots,j_{z-1}) \in g/\{j\}} \prod_{z'=1}^{z-1} \frac{\exp(V_{j_{z'} p})}{\sum_{z''=z'}^{N_{gp}} \exp(V_{j_{z''} p})} \cdot \frac{\exp(V_{jp})}{\sum_{z'''=z}^{N_{gp}} \exp(V_{j_{z'''} p})}$$

where $N_{gp} \subset N_p$ is the set of funds that p considers in category g.

Moving backward in sponsor p's menu choice problem, the probability that n options are chosen from investment category g is given by $q(1-q)^{n-1}$ so that, conditional on g being offered, the probability of j being included in p's plan is just

$$\sum_{n=1}^{\infty} q(1-q)^{n-1} \phi_j^{1:n}.$$

The choice of whether or not to include category g is assumed to depend on sponsors' expected utility from the highest ranked option, which under our T1EV assumption, is given by

$$\mathbb{E}[u_{j_1 p}] = \log \left(\sum_{k \in g} \exp(V_{kp}(\boldsymbol{\theta}_p)) \right).$$

The probability that p decides to offer investment category g as part of its retirement plan

equals

$$\lambda_{gp} = \frac{\exp(\mathbb{E}[u_{j_1 p}])}{1 + \exp(\mathbb{E}[u_{j_1 p}])}$$

which can be interpreted as the probabilistic outcome of a binary choice logit problem.

Combining all pieces together, it can be shown that the unconditional probability of fund j being included in sponsor p plan can be written as

$$\phi_{jp}(\boldsymbol{\theta}_p) = \lambda_{gp}(\boldsymbol{\theta}_p) \cdot \sum_{n=1}^{\infty} (1-q)^{n-1} \phi_{jp}^{n}(\boldsymbol{\theta}_p) \tag{2.6}$$

where I make explicit its dependence on the vector of preference parameters $\boldsymbol{\theta}_p$, $(1-q)^{n-1}$ is the probability that p includes a number of options greater or equal than n and ϕ_j^n is the probability that j is ranked in the nth position.[28]

Equation (2.6) extends the expression of the logit choice probabilities to the case in which decision makers can select more than a single option and where the number of option chosen is determined by the parameter q. Expression (2.6) collapses to the standard discrete choice logit probability if we assume sponsors can only include one fund within each investment category. This corresponds to setting $q = 1$, which implies that

$$\phi_{jp}(\boldsymbol{\theta}_p) = \lambda_{gp}(\boldsymbol{\theta}_p) \cdot \phi_{jp}^{1}(\boldsymbol{\theta}_p) = \frac{\exp(V_{jp}(\boldsymbol{\theta}_p))}{1 + \sum_{k \in g} \exp(V_{kp}(\boldsymbol{\theta}_p))}.$$

In this more general context, the decision of not including investment category g plays the role of the outside option in standard discrete choice models, with mean indirect utility normalized to 0.

Overall, equation (2.6) represents sponsor p individual demand for investment funds belonging to investment category g. Assuming sponsors preferences $\boldsymbol{\theta}_p$ follow some distribution

28. I provide details on the derivation in Appendix 2.11

F_θ, we can derive fund j's aggregate demand as

$$\phi_j = \int \phi_{jp}(\boldsymbol{\theta}_p) dF_\theta(\boldsymbol{\theta}_p). \tag{2.7}$$

The data counterpart to (2.7) corresponds to the share of plans that include fund j as part of their retirement menu. In Section 2.6 I estimate the distribution of sponsors preference parameters by matching (2.7) to these observed inclusion probabilities.

2.4.3 Investors' retirement portfolio problem

Consider investor i who allocates its dollar contribution A across the investment funds available in plan p, indexed by $j \in \{1, ..., J_p\}$, and a cash account $j = 0$. In practice, not all plan investors make an active investment decision and many of them are automatically defaulted into one of the options available which often corresponds to a TDF fund or a balanced fund. I denote this default option $j = d$ and assume that with probability δ investor i does not make an active investment decision. In this case, i's contribution will be allocated entirely to fund d. Conversely, with probability $(1 - \delta)$ investor i makes an active investment decision and allocates A across all options available including d.

Conditional on making an active investment decision, investor i forms its retirement portfolio by choosing the vector of portfolio weights $\boldsymbol{a} \equiv (a_1, ..., a_{J_p})$ to maximize the following linear-quadratic utility:[29]

$$U_p(\boldsymbol{a}) \equiv \underbrace{\sum_{j=1}^{J_p} a_j(\boldsymbol{w}_j'\boldsymbol{\beta} - f_j + \xi_j)}_{\text{linear utility component}} - \underbrace{\frac{\gamma}{2}\sum_{j=1}^{J_p} a_j^2}_{\text{naive diversification}} - \underbrace{\frac{\gamma}{2}\sum_{j,k} g_{jk}a_j a_k}_{\text{funds substitutability}} . \tag{2.8}$$

The preferences defined in (2.8) capture the idea that investors value funds along three

29. In equation (2.8) I have already substituted for the portfolio share on the cash account a_0 using the constraint that $a_0 + \boldsymbol{a}'\mathbf{1} = 1$ and assuming that returns on the cash account are normalized to 0.

margins. The first is a 'perfect substitute' margin that pushes them to allocate their entire contribution to the fund with the highest linear utility component. This component depends on observed funds' characteristics $\boldsymbol{w}_j$, fees f_j, and on unobserved characteristics ξ_j. If this were the only margin, investors' portfolio problem would collapse to a standard discrete choice problem with investors' indirect utility for fund j given by $\boldsymbol{w}_j'\beta - f_j + \xi_j$.

In the context of investment choices, standard portfolio theory predicts that investors should diversify across assets to reduce risk (Markowitz (1952)). The second margin captures this incentive. Specifically, it captures investors' incentives to naively diversify across the available options which, in the context of retirement investing, has been shown to be a key determinant of individual portfolio allocations (Benartzi and Thaler (2001), Huberman and Jiang (2006)).

The third margin captures how investors perceive substitutability between funds. I assume that investors perceive fund j and fund k as substitutable if they belong to the same investment category, in which case the term $g_{jk} = 1$. If funds belong to different investment categories $g_{jk} = 0$. This implies that investors will have an incentive to diversify more across funds from different categories rather than across funds within the same category. As I explain in more detail later on, this assumption is motivated by the empirical evidence that investment categories fixed effects explain a substantial fraction of the observed plan-level portfolio allocation, suggesting that investors' allocate across styles rather than among individual funds (Barberis and Shleifer (2003)). Conversely, funds' loadings on standard risk factors (Fama and French (1992), Carhart (1997)) have negligible explanatory power, especially after controlling for investment categories fixed effects.

2.4.4 Plan asset demand system

Letting G_p be the $J_p \times J_p$ matrix whose (j,k) element is g_{jk}, under the previous assumptions, investor i's optimal portfolio allocation is given by

$$\boldsymbol{a}_i(\boldsymbol{f}_p) = \begin{cases} \boldsymbol{e}_d & \text{if } i \text{ inactive} \\ \frac{1}{\gamma}(I + G_p)^{-1}(W_p\boldsymbol{\beta} + \boldsymbol{\xi}_p - \boldsymbol{f}_p) & \text{if } i \text{ active} \end{cases} \tag{2.9}$$

where $\boldsymbol{e}_d$ is a unit vector that takes value of 1 in its d element corresponding to the default option and W_p is the matrix of observed characteristics for the funds available in plan p.

In the data I do not observe asset allocations at the individual level. Therefore, I need to aggregate individual investors demands to obtain the plan level demand system. Letting $\boldsymbol{s}_p$ be the J_p vector of plan p portfolio shares, A_p plan p total wealth and defining $\boldsymbol{\eta} \equiv (\boldsymbol{\beta}, \gamma, \delta)$ as the vector of demand parameters, we can sum demands across all investors in plan p to obtain:

$$\boldsymbol{s}_p(\boldsymbol{f}; \boldsymbol{\eta}) \equiv \sum_{i \in I_p} \frac{A}{A_p} \boldsymbol{a}_i(\boldsymbol{f}) = \delta \boldsymbol{e}_d + \frac{1-\delta}{\gamma}(I + G_p)^{-1}(W_p\boldsymbol{\beta} + \boldsymbol{\xi}_p - \boldsymbol{f}_p). \tag{2.10}$$

Empirically, I can use variation in the observed plan level allocations to estimate the demand system in (2.10). To this end, it is useful to multiply both sides of (2.10) by $(I + G_p)$ to obtain a demand system where only own fees and own demand shocks enter each equation:

$$\tilde{s}_p(\boldsymbol{f}; \boldsymbol{\eta}) = \delta \tilde{\boldsymbol{e}}_d + W_p\tilde{\boldsymbol{\beta}} - \tilde{\gamma}\boldsymbol{f}_p + \tilde{\boldsymbol{\xi}}_p. \tag{2.11}$$

where $\tilde{\boldsymbol{s}}_p \equiv (I + G_p)\boldsymbol{s}_p$, $\tilde{\boldsymbol{e}}_d \equiv (I + G_p)\boldsymbol{e}_d$, $\tilde{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}(1-\delta)/\gamma$ and $\tilde{\boldsymbol{\xi}} \equiv \boldsymbol{\xi}(1-\delta)/\gamma$. Because $\tilde{\boldsymbol{s}}_p$ is observed, I can estimate (2.11) via linear methods.

Before turning to the supply side, a couple of remarks are in order. First, so far I have assumed that investors' preferences are homogeneous. In this way, the parameters of

the plan-level demand system are the same as the ones for the individual demand system. From an empirical perspective, the nature of the data and the linear structure of the demand system prevent me from learning about unobserved heterogeneity in preference parameters.[30] Nevertheless, in Appendix 2.11, I show how to interpret the parameters in (2.10) as weighted averages of the heterogeneous individual parameters. Second, I am assuming that the default option is the same for each individual. This is done only for expositional purposes and in the estimation and in Appendix 2.11 I allow for the presence of multiple default funds.

Lastly, in setting the supply side profit maximization problem, I will be working with a version of equation (2.10) that I rearrange slightly as follows:

$$\boldsymbol{s}_p(\boldsymbol{f};\boldsymbol{\eta}) = \delta \boldsymbol{e}_d + \frac{(1-\delta)}{\gamma}(I - \mathcal{K}_p)(\boldsymbol{\mu}_p - \boldsymbol{f}) \tag{2.12}$$

where $\boldsymbol{\mu}_p \equiv W_p\boldsymbol{\beta} + \boldsymbol{\xi}_p$ and the matrix $\mathcal{K}_p$ is defined as

$$\mathcal{K}_p \equiv \tilde{G}_p(I + \tilde{G}_p'\tilde{G}_p)^{-1}\tilde{G}_p',$$

with $\tilde{G}_p$ a matrix with J_p rows, one for each fund in plan p, and a number of columns equal to the number of investment categories. The jth row of $\tilde{G}$, denoted $\tilde{\boldsymbol{g}}_j$, equals 1 in correspondence of fund j's category and 0 otherwise. The matrix G_p that appears in (2.10) is the outer product of $\tilde{G}_p$, e.g., $G_p = \tilde{G}_p\tilde{G}_p'$ with $g_{jk} = \boldsymbol{g}_j'\boldsymbol{g}_k$.

Rewriting aggregate asset demand as in (2.12) helps in visualizing own and cross substitution patterns across different assets. In particular, the fee elasticity between asset j and

30. Recall that I do not observe portfolio allocations at the individual level. One way to introduce heterogeneity would be adding interaction terms between funds' characteristics, such as fees, and observable plan characteristics. Egan, MacKay and Yang (2023) take this approach to uncover heterogeneity in investors' risk aversion.

asset l is proportional to

$$\frac{\partial s_j}{\partial f_l} \propto \begin{cases} -(1-\kappa_{jj}) = -\left(1 - \tilde{\boldsymbol{g}}_j'(I + \tilde{G}_p\tilde{G}_p')^{-1}\tilde{\boldsymbol{g}}_j\right) & \text{if } j = l \\ \kappa_{jk} = \tilde{\boldsymbol{g}}_j'(I + \tilde{G}_p\tilde{G}_p')^{-1}\tilde{\boldsymbol{g}}_l & \text{if } j \neq l \end{cases} \tag{2.13}$$

which is always between $(-1, 1)$ if $j \neq k$ and between $(-1, 0)$ if $j = l$. If one defines $\tilde{G}$ as the matrix of funds' factor loadings, $\frac{\partial s_j}{\partial f_l}$ measures how close/correlated asset j and l are in terms of their risk exposures $\tilde{\boldsymbol{g}}_j$ and $\tilde{\boldsymbol{g}}_l$ respectively.[31] When $\tilde{G}$ is the matrix of funds categories fixed effects the same interpretation applies but the substitution patterns are by construction sparse.

2.5 Supply

I model supply as a differentiated Bertrand oligopoly where investment advisors set fees simultaneously before sponsors make their plan menu decisions and plan investors form their portfolios. I assume that the same fund charges the same fee across different plans because, although funds can price discriminate by offering different share classes, almost all the observed variation in fees is across funds and not across share classes within the same fund.[32]

Let P be the number of plan sponsors, A_p the dollar wealth of plan p and $\mathcal{S}_{jp} \subseteq 2^{N_p}$ be the set of all possible menus where p includes fund j. Fund j chooses its fee f_j to maximize the following expected dollar profit

$$\max_{f_j} \quad P \cdot (f_j - c_j) \cdot \int_p A_p \left(\sum_{S \in \mathcal{S}_{jp}} \phi_p(S, \boldsymbol{f}; \boldsymbol{\theta}_p) s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p | S) \right) dF(A_p, \boldsymbol{\theta}_p, \boldsymbol{\eta}_p) \tag{2.14}$$

31. I provide more details in Appendix 2.11

32. Additionally, in the context I am considering, 401(k) sponsors are almost always treated as institutional investors and many investment providers offer a specific share class for the retirement plans market.

where $\phi_p(S, \boldsymbol{f}; \boldsymbol{\theta}_p)$ is the probability that sponsor p chooses plan menu S and, conditional on menu S, $s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p|S)$ is fund j's portfolio weight within plan p

$$s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p|S) = \begin{cases} \frac{1-\delta_p}{\gamma_p}\left((1-\kappa_{jj}^S)(\mu_{jp}-f_j) - \sum_{l\neq j, l\in S}\kappa_{jl}^S(\mu_{lp}-f_l)\right) & \text{if } j \neq d \\ \delta_p + \frac{1-\delta_p}{\gamma_p}\left((1-\kappa_{dd}^S)(\mu_{jp}-f_j) - \sum_{l\neq j, l\in S}\kappa_{dl}^S(\mu_{lp}-f_l)\right) & \text{if } j = d \end{cases}$$

where I made explicit the dependence of the elements of $\mathcal{K}$ on the realized menu S and allow investor parameters to depend on p.

Problem (2.14) is particularly complex to solve because it requires investment providers to internalize how a marginal increase in fees affects (i) investors portfolio allocation s_j conditional on a given menu S, (ii) the probability that plan menu S is chosen and (iii) trade-off these changes across all possible plan menus S and plan sponsors p. From a computational perspective, given the large number of funds available in the market, the number of possible plan menus in $\mathcal{S}_{jp}$ would be too large to make the computation of (2.14) and its derivatives feasible.[33] To overcome these difficulties, I will simplify funds' pricing problem in a way that allows me make the problem computationally tractable while at the same time preserving the two dimensions along which funds' compete in the retirement market, namely, competition for plan inclusion and competition for plan asset allocations.

To simplify problem (2.14) I will assume that funds only consider the effect of a marginal change in fees on their aggregate inclusion probability and not on the probability of any of the possible menus being selected by the sponsor. Equivalently, I assume funds do not take into account that changing fees influences the probability that a particular menu is

33. Goeree (2008) faces a similar problem when estimating a discrete choice demand model with imperfect consideration. She overcomes the computational burden by simulating consumers consideration sets. My case is more complex because investors' consideration sets, or equivalently plan menus, are the outcome of sponsors' menu choice problem. Moreover, consideration probabilities in my case are not independent for products that belong to the same investment category. Lastly, in my case prices affect consideration probabilities which means that derivatives of consideration probabilities will enter firms' first order conditions. This will be true for both own consideration probabilities but also competitors consideration probabilities.

chosen but only consider how it affects their total likelihood of being included. With this assumption, fund j fee setting problem can be rewritten more compactly as

$$\max_{f_j} P \cdot (f_j - c_j) \cdot \int A_p \phi_{jp}(\boldsymbol{f}; \boldsymbol{\theta}_p) s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p) dF(A_p, \boldsymbol{\theta}_p, \boldsymbol{\eta}_p) \tag{2.15}$$

where $\phi_{jp}(\boldsymbol{f}; \theta_p)$ is the probability that p includes j as defined in (2.6) and $s_j(\boldsymbol{f}; \boldsymbol{\eta}_p)$ is the expected portfolio allocation of fund j within plan p and is given by

$$s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p) = \begin{cases} \frac{1-\delta_p}{\gamma_p} \left((1 - \bar{\kappa}^p_{jj})(\mu_{jp} - f_j) - \sum_{l \neq j} \bar{\kappa}^p_{jl} (\mu_{lp} - f_l) \right) & \text{if } j \neq d \\ \delta_p + \frac{1-\delta_p}{\gamma_p} \left((1 - \bar{\kappa}^p_{dd})(\mu_{jp} - f_j) - \sum_{l \neq j} \bar{\kappa}^p_{dl} (\mu_{lp} - f_l) \right) & \text{if } j = d \end{cases} \tag{2.16}$$

with,

$$\bar{\kappa}^p_{jl} \equiv \begin{cases} \sum_{S \in \mathcal{S}_{jp} \cap \mathcal{S}_{lp}} \frac{\phi_p(S, \boldsymbol{f}; \theta_p)}{\phi_{jp}(\boldsymbol{f}; \theta_p)} \kappa^S_{jl} = \phi_{lp}(\boldsymbol{f}, \theta_p) \cdot \mathbb{E}[\kappa^S_{jl} | j, l \in S] & \text{if } j \neq l \\ \sum_{S \in \mathcal{S}_{jp}} \frac{\phi_p(S, \boldsymbol{f}; \theta_p)}{\phi_{jp}(\boldsymbol{f}; \theta_p)} \kappa^S_{jj} = \mathbb{E}[\kappa^S_{jj} | j \in S] & \text{if } j = l \end{cases} \tag{2.17}$$

Problem (2.15) can be obtained from (2.14) after dividing and multiplying the term in the round brackets by ϕ_{jp} and exploiting the linear structure of s_{jp} to rewrite the expectation over S more compactly. The resulting term collapses to $s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p)$ as defined in (2.16), which represents the asset demand from plan p investors that fund j expects before sponsor p chooses its plan menu. Asset characteristics affect this expected demand through the matrix $\bar{\mathcal{K}}^p$, whose (j, l) element, defined in (2.17), captures how much competitive pressure j expects from competitor l. The latter depends on how likely is fund l to be included in plan p and, conditional on that, on how close substitute asset j and asset l are.[34]

My restriction on funds' fee-setting behavior assumes that funds do not internalize how

34. This measure of substitutatibility is given by the second term in (2.17) $\mathbb{E}[\kappa^S_{jl} | j, l \in S; \boldsymbol{f}]$. This is an expectation because κ^S_{jl} depends on the whole menu S and not only on fund j and fund l characteristics. Formally, this can be seen from the definition of κ_{jl} in (2.13) where κ_{jl} depends on the characteristics of all competitors through the weighting matrix $(I + \tilde{G}'_p \tilde{G}_p)^{-1}$.

fees affect the elements $\bar{\kappa}_{jl}$. A constructive way to impose this restriction could be assuming that funds believe sponsors will include at most one fund per investment category, thereby assigning positive probability only to plan menus containing funds from different categories. Formally, this would require funds to have a biased belief $\hat{q} = 1$ about the parameter q governing the distribution of the number of funds included within each category. In practice, we know that most plans do not include more than one option per category (Figure 2.9), making this assumption perhaps not so unreasonable. Under this assumption, I show in Appendix 2.11 that $\bar{\kappa}_{jl}$ would not depend on f_j because plan inclusion decision are assumed to be independent across investment categories. Moreover, this assumption is one of the sufficient conditions that allows me to prove existence of a Bertrand-Nash equilibrium.[35]

2.5.1 Nash equilibrium fees

In this section I derive the Nash equilibrium fees implied by problem (2.15) assuming that funds take $\bar{\mathcal{K}}$ as given. Denoting by $s_j(\boldsymbol{f})$ fund j expected dollar asset allocation, the first order conditions with respect to f_j is given by the usual Bertrand pricing equation

$$s_j(\boldsymbol{f}) + (f_j - c_j) \cdot \frac{\partial s_j(\boldsymbol{f})}{\partial f_j} = 0. \tag{2.18}$$

The difference between the current setting and standard oligopolistic problems is that funds are competing along two dimensions, namely, they compete for being included in a plan and, conditional on plan inclusion, they compete for plan investors' allocations. These two layers of competition are enclosed in $\partial s_j(\boldsymbol{f})/\partial f_j$ which is made of the following two

35. In Appendix 2.11 I show that when $\hat{q} = 1$, sponsors preferences are homogeneous and a particular dominance diagonal condition on the jacobian of the demand system is satisfied there exists a Bertrand-Nash equilibrium.

terms

$$\frac{\partial s_j(\boldsymbol{f})}{\partial f_j} = \underbrace{\int A_p \frac{\partial \phi_{jp}}{\partial f_j}(\boldsymbol{f};\boldsymbol{\theta}_p) s_{jp}(\boldsymbol{f};\boldsymbol{\eta}_p) dF(A_p,\boldsymbol{\theta}_p,\boldsymbol{\eta}_p)}_{\text{demand loss from marginal sponsor}}$$
$$\underbrace{-\int A_p(1-\delta_p)\gamma_p^{-1}\phi_{jp}(\boldsymbol{f};\boldsymbol{\theta}_p)(1-\bar{\kappa}^p_{jj})dF(A_p,\boldsymbol{\theta}_p,\boldsymbol{\eta}_p)}_{\text{demand loss from marginal investor}} < 0. \tag{2.19}$$

Expression (2.19) captures the classic expected revenue loss from marginal consumers not willing to buy at higher price. In this context, the reduction in demand comes from two forces (i) the marginal sponsor not willing to include j in its plan and (ii) the marginal plan investors reducing its allocation to fund j. Equation (2.18) then tells us that, for given competitors fees, fund j will choose f_j that equalizes the profit reduction from losing the marginal sponsors and investors to the profit gain from charging the inframarginal ones an higher fee.

The linear structure of investors asset demand allows me to go beyond this standard intuition and to offer a novel characterization of the Nash equilibrium fees that sheds light on the forces driving price competition in this market. To this end, I will define the following variables,

$$\bar{\phi}_j \equiv \int A_p(1-\delta_p)\gamma_p^{-1}\phi_{jp}dF_p; \qquad \tilde{\kappa}_{jl} \equiv \bar{\phi}_j^{-1}\int A_p(1-\delta_p)\gamma_p^{-1}\phi_{jp}\bar{\kappa}^p_{jl}dF_p$$
$$\bar{\mu}_j \equiv \bar{\phi}_j^{-1}\int A_p(1-\delta_p)\gamma_p^{-1}\phi_{jp}[I-\bar{\mathcal{K}}^p]'_j\boldsymbol{\mu}_p dF_p; \qquad \tilde{\boldsymbol{\mu}} \equiv (I-\tilde{\mathcal{K}})^{-1}\bar{\boldsymbol{\mu}}$$
$$\bar{\delta} \equiv \bar{\phi}_j^{-1}\int A_p\phi_{jp}\delta_p dF_p; \qquad \iota_j \equiv -\bar{\phi}_j^{-1}\int A_p\frac{\partial\phi_{jp}}{\partial f_j}s_{jp}(f_j-c_j)dF_p$$

where I suppressed all functions' arguments and $[I-\bar{\mathcal{K}}^p]_j$ denotes the jth row of the matrix $I-\bar{\mathcal{K}}^p$. Next, I rewrite (2.18) in vector form for all funds in terms of these variables to

obtain

$$\bar{\delta}\boldsymbol{e}_d + (I - \tilde{\mathcal{K}})(\tilde{\boldsymbol{\mu}} - \boldsymbol{f}) - \boldsymbol{\iota} - (I - diag(\tilde{\mathcal{K}}))(\boldsymbol{f} - \boldsymbol{c}) = 0$$

By rearranging this system of Bertrand FOCs, I reach one of the paper's key findings, which decomposes equilibrium fees into three components:

$$\boldsymbol{f}^* = \underbrace{\frac{\tilde{\boldsymbol{\mu}} + \boldsymbol{c}}{2}}_{\text{monopolist fee}} - \underbrace{\boldsymbol{h}\left(\tilde{\mathcal{K}}, \frac{\tilde{\boldsymbol{\mu}} - \boldsymbol{c}}{2}\right)}_{\text{Hotelling markdown}} - \underbrace{\left(I - \frac{diag(\tilde{\mathcal{K}})}{2} - \frac{\tilde{\mathcal{K}}}{2}\right)^{-1} \frac{\boldsymbol{\iota}}{2}}_{\text{plan inclusion markdown}} \tag{2.20}$$

where for simplicity I assumed that there is no default fund.[36]

Equation (2.20) decomposes the fees charged by funds in an interior Nash equilibrium into three terms. The first term represents the vector of fees funds would charge as monopolist.[37] From these fees there are two types of markdowns that need to be subtracted to account for the two dimension of competition driving pricing incentives in this market. The plan inclusion markdown in (2.20) captures the optimal reduction in fees required to increase the probability of being included in a plan. If funds knew that they will be included with certainty i.e., $\phi_{jp} \equiv 1$, the plan inclusion markdown would disappear because $\iota_j = 0$. The Hotelling markdown $\boldsymbol{h}(\cdot)$ instead captures the optimal reduction in fees required to compete against similar funds. The simple Hotelling (1929) model predicts that when two firms located on a line are closer to each other, they will charge lower margins. The same intuition carries over in this more general setting. Funds that are closer to their competitors have an higher $\boldsymbol{h}$ an must lower their fee.

The natural question at this point is, what does being closer to competitors mean in this context and how does that relate to $\boldsymbol{h}$? Loosely speaking a fund is closer to its competitors

36. All derivations are presented in Appendix 2.11

37. The price charged by a monopolist facing a linear demand $q(p) = a - p$ with marginal cost c, is $(a+c)/2$.

when its characteristics are less differentiated from competitors' characteristics. In practice, this measure of proximity/differentiation is embedded in the asset demand cross-substitution patterns through the matrix $\mathcal{K}$ defined in (2.13). Fund j and fund l are closer substitutes if their characteristics are closer as measured by κ_{jl}. The proximity of each fund to all other competitors is summarized by the vector $\boldsymbol{h}(\cdot)$ which is defined as

$$\boldsymbol{h}\left(\tilde{\mathcal{K}}, \frac{\tilde{\boldsymbol{\mu}} - \boldsymbol{c}}{2}\right) \equiv \left(I - \frac{(\tilde{\mathcal{K}} - \kappa_0 I)}{2(1-\kappa_0)}\right)^{-1} \frac{(\tilde{\mathcal{K}} - \kappa_0 I)}{2(1-\kappa_0)} \frac{\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{c}}}{2} \tag{2.21}$$

where for expositional purposes I assumed that $\tilde{\kappa}_{jj} \equiv \kappa_0$. Expression (2.21) is a measure of proximity because it is equivalent to the Bonacich network centrality measure studied in Bonacich (1987).[38] This network centrality measure appears in (2.20) because the Bertrand fee-setting game belongs to a broader class of network games first studied in Ballester et al. (2006).[39] The main insight from this literature is that Nash equilibrium actions will generally depend on a player's network centrality. In this case, funds' equilibrium fees depend on how central a fund is in the competitive network. A more central fund faces more similar competitors and charges lower margins.

So far, I have assumed that a Nash equilibrium exists. Given the complexity of funds' pricing problem deriving general results about existence and uniqueness is not an easy task. Nonetheless, in Appendix 2.11 I provide sufficient conditions for existence and uniqueness of an interior Bertrand-Nash equilibrium for some particular cases. Specifically, I start by showing that when funds know with certainty which plan menus will include them (e.g.,

38. For any zero-diagonal adjacency matrix A, positive scalar $\delta > 0$ and non-zero vector $\boldsymbol{u}$, the vector of Bonacich centralities is defined as $\boldsymbol{b}(A, \boldsymbol{u}) \equiv (I - \delta A)^{-1} \delta A \boldsymbol{u}$.

39. More details are provided in Appendix **??**.

$\phi_{jp} = 1$ if p includes j) then the following dominance diagonal condition

$$(1 - \tilde{\kappa}_{jj})(\tilde{\mu}_j - c_j) > \sum_{l \neq j} |\tilde{\kappa}_{jl}|(\tilde{\mu}_j - c_l) \quad \text{all } j,$$

ensures that the Bertrand-Nash equilibrium is interior, with $f_j^* \in (c_j, \tilde{\mu}_j)$ for all j, and unique. Moreover, I am also able to show that when sponsors' preference are homogeneous and funds believe that sponsors will include at most one fund per category (e.g., $\hat{q} = 1$) there exists a Nash-equilibrium even when funds' do not know with certainty which plan menus will include them.

2.6 Demand Identification and Estimation

In this section I describe how to identify and estimate the model. I estimate sponsors' preferences from variation in the observed plan inclusion probabilities and investors' preferences from variation in the observed plan-level portfolio allocations. After that, I turn to the supply side and recover funds' marginal costs and markups using the demand estimates together with the Nash-Bertrand equilibrium conditions.

2.6.1 Identification and estimation of sponsors' preferences

The menu choice model developed in Section 2.4 provides us with an analytical expression for the expected probability that fund j is included in a retirement plan for a given distribution of sponsors' preference parameters $\boldsymbol{\theta}_p \sim F(\boldsymbol{\theta}_p; \bar{\boldsymbol{\theta}})$ which I assume to be parametrized by the vector $\bar{\boldsymbol{\theta}}$:

$$\phi_j(\bar{\boldsymbol{\theta}}) = \int \lambda_{gp}(\boldsymbol{\theta}_p) \cdot \sum_{n=1}^{\infty} (1-q)^{n-1} \phi_{jp}^n(\boldsymbol{\theta}_p) dF(\boldsymbol{\theta}_p; \bar{\boldsymbol{\theta}}), \tag{2.22}$$

where $\lambda_{gp}(\boldsymbol{\theta}_p)$ is the probability that plan p offers category g and $\phi^n_{jp}(\boldsymbol{\theta}_p)$ is the probability that j is the fund with the nth highest utility.

The data counterpart to equation (2.22) is the share of retirement plans that include fund j. The estimation strategy is then to find the vector of parameters $\bar{\boldsymbol{\theta}}$ that makes the model implied inclusion probabilities in (2.22) as close as possible to the observed ones. As $q \to 1$, expression (2.22) collapses to the standard random-coefficient logit formula for product market shares

$$\phi_j(\bar{\boldsymbol{\theta}}) = \int \frac{\exp(V_{jp}(\boldsymbol{\theta}_p))}{1 + \sum_{k \in g} \exp(V_{kp}(\boldsymbol{\theta}_p))} dF(\boldsymbol{\theta}_p; \bar{\boldsymbol{\theta}}),$$

considered in the workhorse demand models of Berry (1994) and Berry, Levinsohn and Pakes (1995).

In estimation, I compute the observed inclusion probability for each fund at the year-recordkeeper-category level, where I define an investment category as the interaction between the standard investment categories and an indicator for passive funds. In this way, an index fund and active fund from say Large-Cap-Growth would be classified into two different categories Large-Cap-Growth-Active and Large-Cap-Growth-Passive. I refer to a particular year-recordkeeper-category combination as a market, indexed by t, and denote the share of retirement plans in market t that offer fund j as $\hat{\phi}_{jt}$.

I allow for the possibility that funds' fees are correlated with sponsors' demand shocks ζ_{jt}. These shocks enter sponsors' mean utilities $V_{jpt}(\boldsymbol{\theta}_p) = \boldsymbol{w}'_{jpt}\boldsymbol{\theta}_p + \zeta_{jt}$ and are observed by market participants, including investment funds, but unobserved to the econometrician. If funds set fees after observing ζ_{jt} or have better information about these demand shocks, demand and supply simultaneity would bias preference parameters estimates. I account for this type of price endogeneity in two ways. First, I exploit the granularity of the data to absorb unobserved heterogeneity in demand along three dimensions: (i) time by including

year fixed effects, (ii) product quality by including funds' brand fixed effects, and (iii) financial characteristics by including investment category fixed effects. Second, I instrument funds' fees with funds' turnover ratios which capture trading-related costs that funds incur when selling and buying securities. The instrument is relevant as long as profit maximizing funds optimally pass these costs to investors through higher fees (Pástor, Stambaugh and Taylor (2020)). The identifying assumption is that the variation in funds' turnover ratio not explained by time, brand, investment category and passive fixed effects enters sponsors' demand only through fees.

A large literature in finance has studied the relationship between funds' turnover and funds' investment performance. The results are mixed: Carhart (1997) finds a negative cross-sectional relationship, Wermers (2000) and Kacperczyk, Sialm and Zheng (2005) find no relationship, and Pástor, Stambaugh and Taylor (2017) find a positive time-series relationship. Regardless of the sign of such relationship, if investors chase performance (Chevalier and Ellison (1997)) and my set of fixed effects does not control for that appropriately, the exclusion restriction might not hold because turnover may correlate with unobserved demand shocks. In Appendix 2.12 I provide supporting evidence for the validity of the instrument (e.g., the residual turnover) by showing that it does not correlate with standard measure of current and future investment performance for the funds in my sample.

To estimate sponsors preferences I use a nested-fixed point algorithm similar to the one developed in Berry et al. (1995).[40] To start with, I assume that the distribution of sponsors preference parameters is normal with mean $\boldsymbol{\mu}_{\bar{\theta}}$ and variance $\Sigma_{\bar{\theta}}$ (e.g., $\bar{\boldsymbol{\theta}} = (\boldsymbol{\mu}_{\bar{\theta}}, \Sigma_{\bar{\theta}})$) and write sponsor p mean utility as

$$V_{jt}(\bar{\boldsymbol{\theta}}) = \bar{v}_{jt} + \boldsymbol{w}'_{jt}\Gamma_{\bar{\theta}}\boldsymbol{\nu}_p$$

40. See Appendix 2.11 for more details.

where $\bar{v}_{jt} \equiv \boldsymbol{w}'_{jt}\boldsymbol{\mu}_{\bar{\theta}} + \zeta_{jt}$ is the homogeneous component of preferences, $\boldsymbol{\nu}_p \sim N(\mathbf{0}, I)$ are random tastes for funds' characteristics and $\Gamma_\theta \Gamma'_\theta = \Sigma_\theta$. The estimation algorithm starts with a guess of $\bar{\boldsymbol{\theta}}$ and then for each market t finds the vector $\bar{\boldsymbol{v}}_t(\bar{\boldsymbol{\theta}})$ that matches observed and model-implied inclusion probabilities:

$$\hat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t(\bar{\boldsymbol{v}}_t(\bar{\boldsymbol{\theta}})).$$

After that, the demand residuals $\boldsymbol{\zeta}_t(\bar{\boldsymbol{\theta}}) = \bar{\boldsymbol{v}}_t(\bar{\boldsymbol{\theta}}) - W_t\boldsymbol{\mu}_{\bar{\theta}}$ are computed for each market. The last step exploits the orthogonality condition between ζ_{jt} and an appropriate vector of instruments $\boldsymbol{Z}_{jt}$, $\mathbb{E}[\zeta_{jt}|\boldsymbol{Z}_{jt}] = 0$ to form the GMM norm

$$\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}})'\boldsymbol{Z}\Omega(\bar{\boldsymbol{\theta}})\boldsymbol{Z}\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}). \tag{2.23}$$

The algorithm keeps searching over $\bar{\boldsymbol{\theta}}$ until (2.23) is minimized.

2.6.2 Estimates of sponsors' preferences

Table 2.2 presents the estimates of sponsors' preference parameters. The first column reports the estimates of the means of the preference distribution and the second column reports the corresponding standard deviations. These estimates minimize the GMM objective (2.23) following the estimation algorithm I discussed in the previous section.

In the estimation of sponsors' preferences I include five characteristics, two of them are continuous and three of them are binary. The two continuous characteristics are funds' expense ratios, measured in basis points (bp.), and funds' returns in the previous year gross of expenses, measured in percentage points (pp.). Because I absorb investment category fixed effects, returns are relative to the average return of funds within the same category.

The three binary characteristics are indicators for whether a fund is affiliated with the sponsor's recordkeeper, for whether a fund is a target date and their interaction. Including an

indicator for affiliated funds allows me to accommodate for the presence of agency frictions whereby sponsors favor the inclusion of funds belonging to their recordkeeper product line. The inclusion of an indicator for TDFs instead captures the possibility that sponsors have a preference for funds that rebalance plan investors allocation automatically as they age. These type of funds have been created specifically for retirement investing and, after the Pension Protection Act of 2006, qualify as default option for plan participants who do not make an active investment decision. Since then, TDFs' market share in the retirement market has been growing substantially and it is reasonable to think that sponsors may have a preference for such funds even if just to comply with current regulations and reduce liability risk.

The parameter estimates reported in the first column of Table 2.2 suggest that sponsors value more whether a fund is affiliated, and especially if it is an affiliated TDF, rather than how cost-efficient such fund is or how it performed relative to its investment category. The preference coefficient for funds' affiliation is large and significant. The preference coefficient on funds' expense ratios is negative and significant but its magnitude is small if compared with the coefficient on funds' affiliation; on average sponsors are willing to pay 44 bp (=0.88/0.02) more in fees for an affiliated fund. On the other hand, plan sponsors do not seem to value funds' returns gross of fees, as the estimated coefficient is close to zero. I also allow for heterogeneity around the mean of sponsors' sensitivity to fees and report the estimated standard deviation in the second column of Table 2.2. Although modest in magnitude, the estimated heterogeneity is significant at conventional significance levels.

To get a better understanding of the estimated magnitudes, I report the median marginal effect of each characteristic on the inclusion probabilities in the third column of Table 2.2. The marginal effect is the unit change in the expected probability of being included in a plan for a unit increase in the corresponding characteristic. For instance, the first number in the third column tells us that, on average, a ten basis points increase in expenses reduces the plan inclusion probability by almost 0.1 percentage points. On the other hand, the

Employers preference parameters			
	Mean	S.D.	Marginal Effect (pp.)
Expense Ratio (bp.)	-0.020	0.002	-0.008
	(0.002)	(0.000)	
Affiliated (dummy)	0.879	-	0.363
	(0.046)		
Target (dummy)	-0.371	-	-0.123
	(0.088)		
Target × Affiliated	0.194	-	0.081
	(0.096)		
Gross returns (pp.)	0.004	-	0.002
	(0.001)		
Median fee elasticity		-1.77	
q (Calibrated)		0.70	
GMM objective (df)		6.74 (1)	

Table 2.2: Two-step GMM estimates of plan sponsor preferences. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points.

marginal effect of being an affiliated fund is almost four times larger. These magnitudes are far from being negligible given that the median inclusion probability in the estimation sample is roughly 0.52%. The marginal effect of a one percent increase in funds' gross returns is instead substantially smaller, which is not surprising given how small its corresponding preference coefficient is.

The bottom part of Table 2.2 presents some additional information including information about sponsors' elasticity to fees. With the model estimates, I compute the elasticity of inclusion probabilities to fees for each fund-market combination and report the median of this distribution in the bottom part of Table 2.2. The latter is around around 1.77, suggesting that sponsors' demand is not too elastic to funds' fees.

As mentioned before, I allow for the possibility that funds' fees are endogenous but treat other characteristics as predetermined and independent of demand shocks.[41] In estimation I instrument for fees using a third order polynomial of funds' turnover ratios. Overall I

41. This assumption is often used in the empirical industrial organization literature where product characteristics are assumed to be determined before demand shocks are realized.

have seven moments, four included characteristics and three instruments function of funds' turnover, to estimate six model parameters, five means and one standard deviation. The resulting GMM objective is of 6.74 and rejects the overidentifying restriction at the 1% level. This may be due to the fact that there does not seem to be too much heterogeneity in sponsors' preferences even though the model allows for it. The estimates for the homogeneous model are indeed similar (Table 2.9) and come with a GMM objective of 4.57 which, although not perfect, it is not rejected at conventional significance levels.

The nature of the data allows me to assess the heterogeneity of the estimates more directly. I do so along several dimensions. First, I split the sample based on plan size as measured by the number of plan participants and find that smaller plans are less responsive to fees and tend to have a stronger preference for affiliated funds than large plans (Table 2.10). This is consistent with smaller sponsors having less bargaining power and being less willing to pay or search for cheaper investment options (Bhattacharya and Illannes (2022)).[42] Second, I split the sample before and after 2014 and find that sponsors have become more elastic to fees over time (Table 2.11). This is consistent with sponsors, as well as plan investors, becoming more attentive to fees in response to regulatory interventions mandating the disclosure of funds' fees and performance (Kronlund, Pool, Sialm and Stefanescu (2021)).

In Table 2.12 I account for sponsors' inertia by including the lagged plan inclusion probability as an additional characteristic that affects sponsors' demand for a given fund. Intuitively, if sponsors' menus are sticky because of switching costs, we would expect funds' inclusion probabilities to persist over time. If inertia is an important driver of menu choices we would also expect past inclusion probabilities to explain a substantial fraction of the cross-sectional variation in the current inclusion probabilities, possibly reducing the importance of other characteristics such as fees and funds' affiliation. The estimates in Table 2.12

42. The fact that smaller sponsors are more likely to include expensive funds could also reflect the presence of fixed costs in plan provision. For instance smaller sponsors might not have the asset base to access the cheapest share classes of a fund. The model accounts for this because sponsors' consideration sets depend on the size group in Table 2.10.

confirm these intuitions. The coefficient on past inclusion probabilities is large, positive and significant, suggesting that sponsors' inertia is an important driver of plan menu choices. Moreover, when accounting for inertia, the estimated sponsors' sensitivity to fees is lower, with an estimated elasticity to fees of about 1.3. The estimated parameters suggest that, on average, sponsors are willing to pay 17 basis point in fees for a 1% increase in the past inclusion probability.

As an additional robustness check, in the second part of Table 2.9, I estimate sponsors' preferences allowing for heterogeneity in the parameter q governing the distribution of the number of options included within each investment category. Specifically, I estimate q at the year-recordkeeper-category level and find substantially similar results. The reason is that the empirical distribution of the number of options included within each category is essentially the same along several dimensions of heterogeneity one might consider. For example, in Figures 2.16 and 2.17 I plot such distribution for small and large plans, measured by the number of plan participants, and before and after the year 2014. In both cases the distribution of the number of options included within an investment category is virtually unchanged. Similarly, in Figure 2.18 I plot the distribution of the number of options included within each category by broad asset classes and find that, except for bond funds, the distribution is almost identical to the one for the full sample.

2.6.3 Identification and estimation of investors' preferences

The identification of investors' preferences follows closely the logic for the identification of sponsors' preferences. The main difference is that the identifying variation comes from variation in the observed portfolio allocations rather than variation in plan inclusion probabilities.

The estimation of plan investors' preferences is different and simpler than the estimation of sponsors' preferences because investors' demand is linear in preference parameters or some

known function of those. To see this recall that plan p portfolio shares are given by

$$\boldsymbol{s}_p(\boldsymbol{f}_p;\boldsymbol{\eta}_p) = \delta \boldsymbol{e}_p + \frac{(1-\delta)}{\gamma}(I+G_p)^{-1}(W_p\boldsymbol{\beta} + \boldsymbol{\xi}_p - \boldsymbol{f}_p) \tag{2.24}$$

which, after multiplying both sides by $I + G_p$, becomes a system of estimating equations whose RHS only depends on own demand shocks and whose LHS is an observed linear transformation of the observed plan-level portfolio allocations

$$\tilde{\boldsymbol{s}}_p(\boldsymbol{f}_p;\boldsymbol{\eta}_p) = \delta \tilde{\boldsymbol{e}}_p + W_p\tilde{\boldsymbol{\beta}} - \tilde{\gamma}\boldsymbol{f}_p + \boldsymbol{\xi}_p \tag{2.25}$$

with $\tilde{\boldsymbol{s}}_p \equiv (I+G_p)\boldsymbol{s}_p$, $\tilde{\boldsymbol{e}}_d \equiv (I+G_p)\boldsymbol{e}_d$, $\tilde{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}(1-\delta)/\gamma$ and $\tilde{\boldsymbol{\xi}} \equiv \boldsymbol{\xi}(1-\delta)/\gamma$. Equation (2.25) can be estimated via linear regression methods.

As before, I allow for the possibility that funds' fees are correlated with plan investors' demand shocks $\boldsymbol{\xi}_p$. If funds make their price-setting decision after observing $\boldsymbol{\xi}_p$, demand and supply simultaneity would bias preference parameters estimates. To account for this type of price endogeneity I follow the same approach I used for the identification of sponsors' preferences. First, I exploit the granularity of the data to absorb unobserved heterogeneity in demand by including time, funds' brand, passive and investment category fixed effects. In this case, because the estimation is at the fund-plan level I am also able to include plan/sponsors' fixed effects to absorb plan-level preference shocks. Second, I instrument funds' fees with funds' turnover ratio which captures trading costs that are pass on to investors through higher fees. Again, the identifying assumption is that variation in funds' turnover ratio not explained by time, brand, category, passive and plan fixed effects enters investors' demand only through fees.[43]

To assess the robustness of my estimates, I also implement an Hausman-type of identification strategy to account for the endogeneity of fees. Specifically, following Egan, MacKay

43. In Appendix 2.12 I provide more details on using funds' turnover as instrument for funds' fees.

and Yang (2023), I instrument the fee charged by any given fund with the average expense ratio charged by the same fund provider in other investment categories. This instrument will be relevant when a providers cost of operating a mutual fund is correlated with its costs of operating its other mutual funds and when these costs are pass on to investors through fees. The instrument will be excluded if investors' residual demand shocks for any given fund (after controlling for the above series of fixed effects) are uncorrelated with the fees charged by the same fund provider on its funds in other investment categories. The resulting estimates are very similar to the ones I obtain when using funds' residual turnover as instrument for fees.

I estimate investors' preferences applying linear IV to equation (2.25) under the assumption that funds' turnover ratios Z_j are mean-independent of investors' demand shocks, formally, I require that $\mathbb{E}[\xi_{jp}|Z_j] = 0$. To compute the LHS of (2.25) I need to specify which asset characteristics form the matrix $G_p = \tilde{G}_p\tilde{G}_p'$. In classic portfolio theory $\tilde{G}_p$ would include characteristics capturing the correlation structure between assets. Perhaps the most natural way to proceed would be to construct $\tilde{G}_p$ after estimating funds' loadings onto some underlying risk factors from the time-series of funds' returns and then, for each plan, compute the variance-covariance matrix of the funds available, G_p.

Although common in practice, I do not follow this approach and instead use funds' classification into investment categories to construct $\tilde{G}_p$. The reason I do so is that funds' loadings on standard factors do not seem to explain the retirement portfolio allocations observed in the data, whereas investment category fixed effects do. Table 2.13 presents the R-squared from regressing observed portfolio allocations on categories and funds' factors loadings. Factors alone explain close to 4% of the observed variation in portfolio shares whereas investment categories fixed effects alone explain more than three times that. More importantly, after absorbing categories fixed effects, factors' R2 drops substantially to 0.1% suggesting that factors explanatory power was just proxing for investment category classifi-

cations. In a world in which investors' allocation decisions depend on assets' factor structure we would expect factor loadings to have some power in explaining the observed portfolio shares.

Based on this evidence, I use investment categories to model how plan investors interpret assets substitution patterns. I do so by creating a three level nesting structure of asset categories. The first level consists into three broad asset classes Equity, Allocation and Bond. Then I create a second level for each of these classes. For instance, funds belonging to the Equity class are further classified into Equity-Large, Equity-Mid, Equity-Small and Equity-International. In the third level, Equity-Large fund are further classified into Equity-Large-Blend, Equity-Large-Growth and Equity-Large-Value and similarly for other second level categories. I consider each of these levels as a separate asset characteristic corresponding to a column of the matrix $\tilde{G}_p$; for instance if j is an Equity-Large-Value fund then the jth row of $\tilde{G}_p$ is a vector $\tilde{\boldsymbol{g}}_j$ that takes value 1 for the Equity, Equity-Large and Equity-Large-Value columns and 0 everywhere else. The outer product of this matrix of category indicators $G_p = \tilde{G}_p \tilde{G}_p'$ is then a three-block diagonal matrix whose element (j, l) element $\boldsymbol{g}_{jl} = \tilde{\boldsymbol{g}}_j' \tilde{\boldsymbol{g}}_l$ equals 3 if fund j and l belong to the same 3rd level category (e.g., both are Equity-Large-Value funds), equals 2 if they belong to the same 2nd level but to a different 3rd level category, equals 1 if they only belong to the same 1st level and equals 0 otherwise.[44]

2.6.4 Estimates of investors' preferences

Table 2.3 presents the estimates of investors' preference parameters based on the linear specification in equation (2.25). The first three columns present some OLS estimates whereas the fourth column reports the IV estimates from instrumenting funds' fees with funds' turnover ratios. The estimates reported correspond to the coefficients on the RHS of equation (2.25).

44. Another way to see this is thinking about G_p as a quadratic interaction of fixed effects where each level of classification is a fixed effect. I provide an illustrative example in Appendix 2.11.

Besides funds' expenses, I assume that past returns gross of fees and funds' affiliation enter the set of asset characteristics W_p determining the linear component of plan investors' preferences $W_p\boldsymbol{\beta} - \boldsymbol{f}_p + \boldsymbol{\xi}_p$. Investment categories also enter this linear component of investors' utility because I absorb category fixed effects in all specifications. If one interprets this linear component as investors' subjective expectations about assets' returns, the implicit assumption is that investors' subjective beliefs depend on funds' past returns relative to their corresponding investment category.

The OLS estimates broadly suggest that plan investors dislike fees, like returns and have a preference for funds' that are affiliated with their sponsor recordkeeper. A closer look at the magnitudes further reveals that plan investors care more about funds' returns than their sponsors because, in this case, the preference coefficient on returns and its corresponding marginal effect are substantially larger.[45] This is consistent with sponsors designing their plan to merely comply with regulation and minimize liability risk. Current ERISA regulation indeed prescribes that sponsors are not liable for funds' market performance to the extent that their plan includes high quality options compared to the alternative available in the market. Because fees are known whereas performance is uncertain, it is not surprising that sponsors care more about fees rather than gross performance as they cannot be held accountable for the latter.

Plan investors seem to value funds' affiliation less than their sponsors. For the latter, funds' affiliation is a crucial driver of plan inclusion decisions whereas for plan investors the importance of funds' affiliation is more modest although not irrelevant. This is consistent with agency frictions mostly biting at the plan design stage where sponsors tend to favor affiliated funds when constructing their retirement plan (Pool, Sialm and Stefanescu (2016)). Agency frictions could potentially spillover to the investment stage if, for instance, record-

45. Nevertheless, investors, like sponsors, still care much more about fees relative to returns. Although the coefficient magnitudes are similar, in Table (2.3) fees are measured in percentage points whereas returns are measured in decimal form.

Plan investors preference parameters					
	OLS			IV	ME
Expense ratio ($\tilde{\gamma}$)	-0.004	-0.011	-0.012	-0.044	-0.098
	(0.001)	(0.001)	(0.001)	(0.004)	
Affiliated ($\tilde{\beta}_1$)	0.001	0.004	0.017	0.015	0.034
	(0.000)	(0.000)	(0.001)	(0.001)	
Gross returns ($\tilde{\beta}_2$)	0.002	0.012	0.013	0.020	0.043
	(0.001)	(0.001)	(0.001)	(0.001)	
Fraction inactive (δ)	0.267	0.290	0.407	0.371	-
	(0.006)	(0.006)	(0.008)	(0.009)	
Median fee elasticity	-0.225	-0.645	-0.688	-2.517	-
Median fee elasticity (active)	-0.306	-0.909	-1.162	-4.002	-
Fstat	-	-	-	82.55	-
R2	0.24	0.25	0.42	0.82	-
Fund brand FE	N	Y	Y	Y	-
Employer FE	N	N	Y	Y	-

Table 2.3: Estimates of plan investors preferences. All specifications include year, category and passive fixed effects. Expense ratios are in percentage points (pp.). R2 for IV columns is first stage. ME are the (median) marginal effects for portfolio allocations in pp. for a basis point increase in expenses or a pp. increase gross returns.

keepers also offered advising services to plan investors and were to push investors towards affiliated funds. Although I am not aware of any empirical evidence on mis-advising for the particular context I am considering, a theoretical literature in financial economics contemplates this possibility (Inderst and Ottaviani (2012a), Inderst and Ottaviani (2012b)).

The coefficient on fees increases moving from the first to the third column of the OLS estimates. In particular, it more than doubles when I control for fund brand fixed effects. This is consistent with fund brands potentially capturing a good amount of unobserved heterogeneity in investors' preferences for particular fund brands thereby making funds of the same investment provider (and within the same category) perceived as more substitutable to each other. The same happens to the coefficient on funds' gross returns which is not surprising as we expect investors to care about returns net of fees. Including plan (or equivalently sponsor) fixed effects does not affect too much fees and returns coefficients but triples the coefficient on funds' affiliation suggesting that the extent with which agency

frictions matter might depend on sponsor-level unobservables such as sponsors bargaining power in negotiations with the recordkeeper (Bhattacharya and Illannes (2022)). Lastly, after controlling for sponsor fixed effects, the model estimates that nearly two out of five investors do not make an active investment decision.

I report the IV estimates in the fourth column of Table 2.3. All coefficients, except for the one on funds' fees, are largely unchanged. Conversely, the coefficient on funds' expenses is almost four times larger. The discrepancy between OLS and IV estimates is common in contexts where prices and quantities are determined simultaneously in equilibrium. OLS estimates often imply inelastic demand curves because the observed variation in quantity and prices is also due to shifts in demand. However, after instrumenting for prices, the resulting estimates recover demand curves that are much more elastic. This pattern emerges clearly when looking at the elasticity to fees of investors' portfolio allocations, which I report in the second part of Table 2.3. Plan investors' portfolio allocations are inelastic under OLS but become elastic after instrumenting funds' fees with funds' turnover ratios. The median elasticity is around 2.5, almost 50% larger than sponsors' elasticity, indicating that sponsors may not be internalizing investors' preferences when constructing their plan menu. The misalignment in elasticities increases even more if we consider investors who actively form their retirement portfolio. The median fee elasticity for active investors is around 4, over twice larger than sponsors' elasticity to fees. The estimates remain largely unchanged if I use Hausman instruments instead of funds' residual turnover ratios (Table 2.14).

To get a better sense of the magnitudes the last column of Table 2.3 reports the marginal change in investors' portfolio allocations implied by a unit change of the corresponding characteristic. The change in allocations is measured in percentage points per basis point increase in expenses and per percentage point increase in gross returns. A 10 basis points increase in fees reduces the corresponding portfolio allocation by nearly 1%. This magnitude is not small if one considers that the average retirement portfolio allocation is of about 3%.

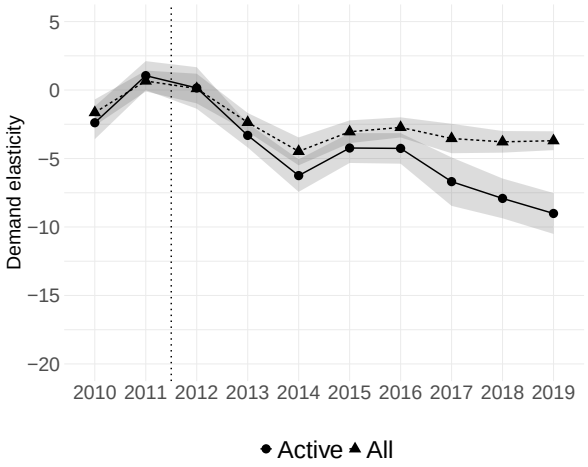

Figure 2.5: Cross-sectional estimates of plan investors median portfolio elasticity to funds' fees. Dashed vertical line corresponds to the DOL fee disclosure reform.

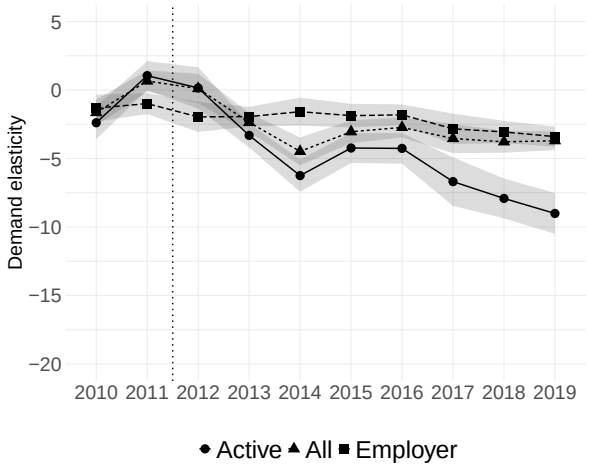

Figure 2.6: Cross-sectional estimates of plan investors median portfolio elasticity to funds' fees. Dark blue squared-dashed line is the median sponsor elasticity.

Conversely, a 1 percentage point increase in past gross returns increases the corresponding portfolio allocation by a modest 0.04%. Because I am absorbing category fixed effects, the latter should be interpreted as the effect of a 1% increase in funds' performance, measured relative to the corresponding investment category, on their plan-level portfolio allocation. The modest magnitude suggests that plan investors do not chase performance as much as documented for the whole mutual fund industry (Chevalier and Ellison (1997)).

The model estimates provided in Table 2.3 pool together all cross-sections from 2010 to 2019, although the identifying variation remains cross-sectional because I always include year fixed effects. That being said, nothing prevents me from estimating investors' preferences separately for each observed cross-section of plan menus and assess how such estimates have changed over time.

Figure 2.5 plots the estimated median fee elasticity for each cross-section from 2010 to 2019. Two broad patterns emerge. First, investors seem to have become more sensitive to fees over time, with a sharp drop in the estimated elasticity from 2011 to 2013. Second, investors have become more inactive, with the estimated elasticity for active investors diverging from

the elasticity of all investors. The first pattern could be a consequence of the regulatory push that required plan sponsors and investment providers to disclose investment fees to plan investors. Specifically, starting from the year 2012 the Department of Labor (DOL) required plan sponsors to disclose information about funds' expenses and performance directly to plan investors and recent empirical evidence suggests that investors have become significantly more attentive to fees as a consequence of that (Kronlund, Pool, Sialm and Stefanescu (2021)). Sponsors' elasticity has also been decreasing over time from around -1.3 in 2010 to about -3.4 in 2019 (Figure 2.6). Except for the years before the DOL reform, sponsors tend to be less elastic to fees than investors, especially if compared to active investors.

The second pattern is likely a symptom of the growth in the demand and supply of TDFs following the 2006 Pension Protection Act which identified TDFs as one of the qualified default investment alternatives for retirement plans. Since then, TDFs have become a constant component of retirement plan menus with more than 80% of sponsors offering at least one TDF in their plan as of 2019 (Figure 2.20). At the same time, plan investors have been increasing their TDFs holdings, with the average portfolio share of TDFs across plans growing three-folds from approximately 10% in 2010 to more than 30% as of 2019 (Figure 2.21).[46] My model attributes this increase in TDFs' portfolio share to the presence of more inactive investors. Indeed, the estimated fraction of inactive investors increases from roughly 25% in 2010 to 60% in 2019 (Figure 2.7), matching closely the share of investors holding a single TDF as reported by Vanguard (2022). This large increase in the share of investors that do not actively form their portfolio explains why the estimated fee elasticity for active investors diverges from the fee elasticity of all investors and why the latter moves closer to sponsors' estimated elasticity (Figure 2.6).

46. Interestingly, this increase in TDFs' market shares has not been accompanied by a reduction in TDFs fees. Relatively to other type of funds TDFs have experienced a much lower decline in fees (Figure 2.22 and 2.23)

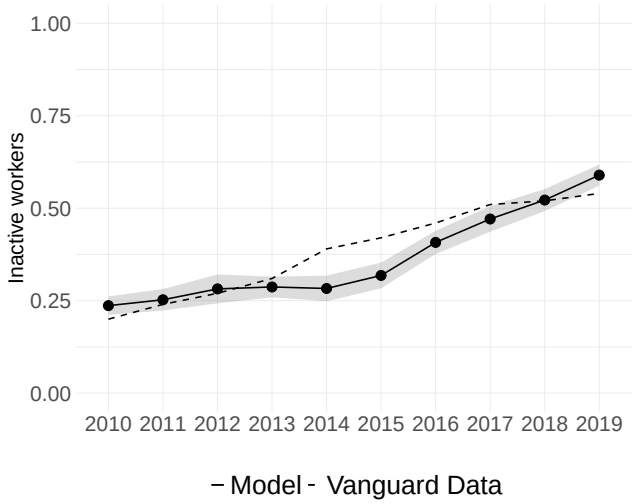

Figure 2.7: Cross-sectional estimates of the fraction of inactive investors (black). Share of plan investors in Vanguard plans holding a single TDF.

2.7 Price-cost Margins and Fee Decomposition

In this section I combine the estimates of sponsors' and investors' preferences together with the Nash-Bertrand first order conditions derived in (2.19) to recover funds' price-cost margins. After that, I exploit the characterization of equilibrium fees derived in equation (2.20) to decompose the observed variation in fees into the monopolist margin, the Hotelling markdown and the plan inclusion markdown.

2.7.1 Recovering funds' price-cost margins

To begin with, I rewrite funds' profit maximization problem making explicit its dependence on the various dimensions of variation I have in the data. I index time periods (i.e., years) by t and recordkeepers by r. Next, I denote by R_{jt} the set of recordkeepers that include fund j in their network of funds and by P_{rt} the set of retirement plans administered by recordkeeper r. I assume that funds set fees simultaneously in each period before sponsors form their retirement menu but knowing which recordkeeper networks they belong to. I

rewrite problem (2.14) as follows:

$$\max_{(f_{jt})_t} \quad \sum_t \sum_{r \in R_{jt}} P_{rt} \cdot (f_{jt} - c_{jt}) \cdot \int \phi^r_{jpt}(\boldsymbol{f}; \boldsymbol{\theta}_p) s^r_{jpt}(\boldsymbol{f}; \boldsymbol{\eta}_p) A_p dF(A_p, \boldsymbol{\theta}_p, \boldsymbol{\eta}_p) \tag{2.26}$$

where I made explicit the dependence of the inclusion probabilities and portfolio shares on the identity of the plan recordkeeper. This dependence is a consequence of the fact that different recordkeepers have different networks of funds.

The first order condition associated with problem (2.26) imply the following price-cost margins for fund j:

$$f_{jt} - c_{jt} = -\frac{\sum_{r \in R_{jt}} P_{rt} \cdot \int \phi^r_{jpt} s^r_{jpt} A_p dF_p}{\left(\sum_{r \in R_{jt}} P_{rt} \cdot \int \frac{\partial \phi^r_{jpt}}{\partial f_{jt}} s^r_{jpt} A_p dF_p\right) - \left(\sum_{r \in R_{jt}} P_{rt} \cdot \int \phi^r_{jpt} (1 - \delta_p) \gamma_p^{-1} (1 - \bar{\kappa}^{r,p,t}_{jj}) A_p dF_p\right)} \tag{2.27}$$

where the two addends in the denominator represent the revenue loss from the marginal sponsor and the revenue loss from the marginal investor respectively.

By plugging in (2.27) the estimated distributions of sponsors' and investors' preference parameters one obtains the margins charged by funds in an interior Nash-Bertrand equilibrium. I report the estimated margins and marginal costs in Table 2.4. The first set of columns presents the estimates for the full sample of funds, whereas the remaining sets of columns focus on active funds, passive funds and TDFs respectively.[47] Starting from the sample of all funds the estimates suggest that the median fund charges a margin of about 14 basis points and a median markup around 20%.

Perhaps not surprisingly things change when looking at passive funds. The median passive fund has a marginal cost more than three times lower than the median fund among all funds and charges a margin that is around 6 basis points. Interestingly, although the absolute margin for the median passive fund is twice smaller than the margin for the median fund

47. I provide more details on the derivation in Appendix 2.11

	All Funds			Active Funds			Passive Funds			Target Date Funds		
	Fee	MC	PCM	Fee	MC	PCM	Fee	MC	PCM	Fee	MC	PCM
p25	43	27	8	63	45	9	9	2	3	13	2	8
p50	74	58	14	85	70	15	22	13	6	32	19	14
p75	101	87	19	109	95	21	45	36	10	52	39	15
Mean	73	59	14	85	70	15	29	21	8	37	26	11

Table 2.4: Price cost margins and marginal costs implied by the Nash-Bertrand first order conditions. Magnitudes are in basis points.

among all funds, in relative terms, it charges a markup of about 30%. This is suggestive of the fact that, although passive funds are typically perceived as more homogeneous products, they still enjoy substantial market power and do not pass all their cost efficiency down to investors (Hortaçsu and Syverson (2004)). Compared to all funds taken together, the distribution of costs for passive funds is more skewed, with the average fund bearing a marginal cost of about 21 basis points. However, the absolute margin charged by the average fund is similar to the median suggesting that the skewness is mostly driven by the cost structure.

The estimated margins and costs for Target Date Funds are reported in the last set of columns of Table 2.4. Starting from the fees, we can see that TDFs tend to be more expensive than passive funds but cheaper than all funds taken together, with the median and average TDFs charging an expense ratio of about 32 and 37 basis points respectively. On the other hand, the estimated cost structure for TDFs is not too dissimilar to the one for passive funds, suggesting that TDFs, although being cost-efficient investment vehicles, charge higher margins to investors. This is particularly evident for the most efficient TDFs. In fact, the margins for the TDFs in the 50th and 25th percentiles are of 14 and 8 basis points respectively, with a median markup of about 39%.

TDFs' pricing power comes from two sources. First, most TDFs are funds affiliated with the plan recordkeeper. Because sponsors value funds' affiliation, inclusion probabilities will be elastic to TDFs' fees. Second, TDFs are the default option in the vast majority of plans,

allowing them to capture assets from inactive investors who do not respond to fees. Recent empirical evidence suggests that TDFs charge excessive fees because they are structured as funds of funds and, as such, their expenses reflect multiple layers of fees. Moreover, the vast majority of a TDF's holdings are funds that belong to same fund family as the TDF itself and, some TDFs tend not to include the cheapest share classes of such funds (Brown and Davies (2021), Sandhya (2011)).[48]

2.7.2 Decomposition of equilibrium fees

In this section I decompose the observed fees exploiting the decomposition derived in equation (2.20). Specifically, I use the estimated preference parameters and the estimated marginal costs to decompose the observed fees into (1) monopolistic fee, (2) Hotelling markdown and (3) plan inclusion markdown. I perform this decomposition for each year of the sample from 2010 to 2019 and obtain the following decomposition for each fund j in each year t

$$f_{jt} = \frac{\tilde{\mu}_{jt} + c_{jt}}{2} - h_{jt} - \bar{\iota}_{jt} \tag{2.28}$$

where $\bar{\iota}_{jt}$ is the jth component of the vector

$$\left(I - \frac{diag(\tilde{\mathcal{K}}_t)}{2} - \frac{\tilde{\mathcal{K}}_t}{2} \right)^{-1} \frac{\boldsymbol{\iota_t}}{2}. \tag{2.29}$$

To get a sense of the magnitudes, Table 2.15 shows the average of each of the three components across all funds and cross-sections. The first column reports the average observed fee which is about 66 basis points.[49] The last three columns instead show the averages for

48. Agency frictions not only impact TDFs fee setting behavior but also their risk-taking incentives as recently documented by Balduzzi and Reuter (2018).

49. This fee is slightly lower than the overall average fee because, to reduce the computational burden, I performed the decomposition only including the 200 largest funds in each recordkeeper network of funds. On average the 200 largest funds accounted for more than 80% of the AUM managed by the recordkeeper.

each of the three components respectively. A monopolist, on average would charge a fee of about 120 basis points but, because of competition, it needs to give up about 44% of such fee. On average, the two competitive markdowns reduce the monopolist fee by nearly 54 basis points. The contribution of each of those is similar, suggesting that competition for entering investors' choice sets and competition in terms of product characteristics are equally important. On average, the Hotelling and inclusion markdowns each erode more than 20% of the fee a monopolist would be able to charge to its consumers.

Figure 2.24 repeats the same exercise for each cross-section from 2010 to 2019. The black solid line represents the average fee and shows its declining trend from around 80 basis points in 2010 to nearly 50 basis points in 2019. The decomposition sheds light on the sources of this decline. The blue bars suggest that the decline in fees is not a consequence of changes in investors willingness to pay, as captured by $\tilde{\mu}$, nor of changes in technological primitives as captured by funds' marginal costs c. The monopolist fee $(\tilde{\mu} + c)/2$ indeed has been roughly stable over time fluctuating between 100 and 130 basis points. On the other hand, the two markdowns seem to be the driver of such declining trend in fees. In absolute terms, they went from reducing the monopolist fee by about 29 basis points in 2010 to nearly 78 basis points in 2019. In relative terms, they accounted for a 27% reduction of the monopolist fee in 2010 which has more than doubled over time, accounting for a 59% reduction in 2019. Overall, these patterns are consistent with both sponsors and investors becoming more sensitive to funds' expenses.

2.8 Counterfactuals

In this section I evaluate the effects of three policy counterfactuals regulating the design of retirement plans. First, I consider the elimination of agency frictions whereby sponsors favor funds affiliated with their plan recordkeeper. Second, I consider the effects of a policy that mandates the inclusion of low-cost options such as low-cost S&P 500 index funds trackers or

low-cost TDFs. Lastly, I consider a policy that caps funds' expenses.

For all counterfactuals, I quantify how the policy in question impacts plan investors welfare and plan expenses relative to the status quo. The latter corresponds to the welfare and expenses computed for the observed plan menus, plan expenses and portfolio allocations. The main takeaway is that mandating the inclusion of low-cost default options and imposing expense ratio caps are the most effective policies. Assuming that sponsors do not value funds' affiliation does not improve investors' outcomes because nothing prevents them to include expensive options that are not affiliated. Similarly, mandating the inclusion of low-cost index funds has only a modest effect on welfare and expenses. The reason is that sponsors will still be including expensive funds and investors will still be investing in those either because they are inactive or for diversification purposes. Mandating the inclusion of low-cost TDFs improves outcomes because inactive investors benefit from having access to cheaper default options.

To measure investors surplus I rely on the quadratic specification of investors' portfolio problem defined in (2.8). At the optimal portfolio allocation, the surplus for active investor i in plan p can be written as,[50]

$$IS_i \equiv \frac{1}{2}\sum_{j=1}^{J_p} a_{ji}(\boldsymbol{f};\boldsymbol{\eta}_p)(\mu_{jp} - f_j)$$

where $\mu_{jp} = \boldsymbol{w}'_{jp}\boldsymbol{\beta} + \xi_{jp}$. Investors' surplus is the sum of the areas below the demand curves of each asset. Because preferences are quadratic and, in turn, the demand for each asset is linear, this area corresponds to a rectangular triangle with height given by $(\mu_{jp} - f_j)$ and base given by given by a_j. Integrating over all active investors we obtain the average surplus

50. See Appendix 2.11 for a derivation.

for a plan p active investor:

$$IS_p^{\text{active}} = \frac{1}{2}\sum_{j=1}^{J_p} s_{jp}^{\text{active}}(\boldsymbol{f};\boldsymbol{\eta}_p)(\mu_{jp} - f_j)$$

where

$$s_{jp}^{\text{active}} \equiv \sum_{i \in I_{p,\text{active}}} \frac{A}{(1-\delta)A_p} a_{ji}.$$

To obtain a complete welfare measure I need to incorporate the surplus of inactive investors. Because, in the model I do not specify any preference for these investors, I define their surplus as

$$IS_p^{\text{inactive}} = \frac{1}{2}(\mu_d - f_d)$$

where d is plan p's default option. The surplus for a plan p investor is given by

$$IS_p = \delta \cdot IS_p^{\text{inactive}} + (1-\delta) \cdot IS_p^{\text{active}}$$

and, the overall surplus is then

$$\int IS_p(\boldsymbol{f};\boldsymbol{\eta}_p) dF(\boldsymbol{\eta}_p). \tag{2.30}$$

Each counterfactual amounts to (1) imposing the policy, (2) solve for funds' counterfactual equilibrium fees, (3) simulate sponsors' plan menus under the counterfactual policy, (4) compute investors' counterfactual portfolios and compute their surplus from equation (2.30) . Because investors have preferences over their portfolio allocations, their surplus is mea-

sured in units of (subjective) excess returns net of fees (e.g., $\mu_{jp} - f_j$). For active investors, I recover μ_{jp} from their estimated preference parameters by computing $\mu_{jp} = \boldsymbol{w}'_{jp}\hat{\boldsymbol{\beta}} + \hat{\xi}_{jp}$, where $\hat{\xi}_{jp}$ are obtained from the residuals of the linear regression in (2.11) multiplied by the estimated $\tilde{\gamma}$. For inactive investors, I use the average annual (excess) return to measure μ_d.

2.8.1 Eliminating preference for affiliated funds

The first counterfactual I consider restricts sponsors' preferences by forcing them not to value funds' affiliation when constructing their plan menu. Under this restriction, holding every other characteristic constant, an affiliated fund and a non-affiliated fund will have the same likelihood of being included in a given plan. A practical way to implement such policy would be issuing penalties for sponsors that are found to be favoring affiliated funds even though they exhibit worse performance than otherwise similar alternatives (Pool, Sialm and Stefanescu (2016)).

The second row of Table 2.16 presents the results from this counterfactual exercise and shows that such policy would be ineffective, leaving investor surplus and plan expenses almost unchanged. In fact, plan expenses decrease by 1 basis point and investor surplus decreases by 3 basis points.

This policy is ineffective because removing sponsors' preference for affiliated funds does not prevent them from including expensive funds that are not affiliated with their record-keeper. For this reason, counterfactual plan expenses remain unchanged as well as sponsors' sensitivity to funds' fees.

Investors' surplus is also almost unaffected. The slight decrease in their surplus could be either because investors have a small preference for affiliated funds or because affiliated funds are not systematically worse than other alternatives. For example, in many cases TDFs are affiliated with the plan recordkeeper and, we know that their expenses are typically below average (Table 2.4) and are particularly valued by inactive investors.

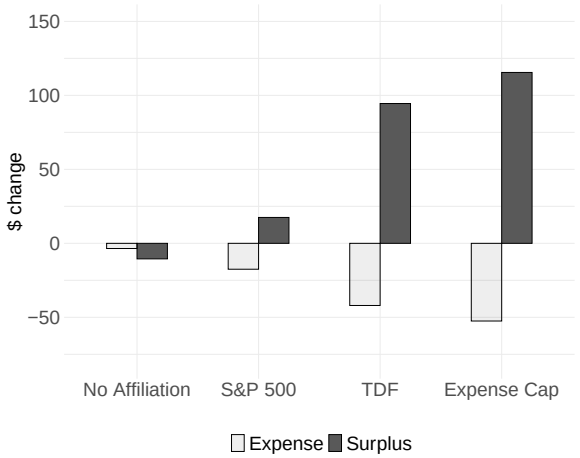

Figure 2.8: Dollar change in surplus and average plan expense relative to the status quo for an investor holding a retirement account of $35,000 under different plan design policies. Magnitudes are in dollars-per-year.

2.8.2 Mandating the inclusion of low-cost options

The second set of counterfactuals studies the effect of policies mandating the inclusion of low-cost investment options. I consider both the inclusion of low-cost index funds that track the S&P 500 and the inclusion of low-cost TDFs. Both policies improve investors' outcome and lead to a reduction of the average plan expenses relative to the status quo.

Mandating the inclusion of at least one low-cost index fund increases investors' surplus by 2% and decreases the average plan expense by 10% relative to the status quo. In magnitudes, the surplus for an investor with a $35,000 account balance increases by about $20 per year (Figure 2.8). To add perspective, in the last column of Table 2.16 I consider the dollar savings over 40 years for an household receiving an annual income of $70,000 and contributing 10% to its 401(k) every year.[51] Such household would save approximately $12,000 in fees after the implementation of this policy.

While one might have expected significant impact from such a policy, its actual effects are relatively modest. This is noteworthy because I selected low-cost funds renowned for being

51. For such computation I assume an annual return of 6%.

both affordable. Each comes with an expense ratio well under 10 basis points, complemented by a 5-star Morningstar rating. My selection includes the Vanguard 500 Index fund (VFIAX) and ETF (VOO), Fidelity 500 Index Fund (FXAIX), Schwab S&P 500 Index Fund (SWPPX), Blackrock iShare Core S&P 500 ETF (IVV), and SPDR S&P 500 ETF (SPY).

Despite this, there are a couple of key reasons why this policy has had only a modest effect on investors' welfare. Firstly, although investors value lower fees, they also want to diversify across all assets available. As such, they will substitute toward the low-cost index only up to the point that does not hurt their diversification needs. This in practice requires maintaining part of the holdings into more expensive funds. Secondly, a significant segment of investors remain inactive in their investment approach. Instead of actively selecting funds, they default their contributions to the available TDF. The addition of a low-cost index fund does not alter the behavior of this group. Furthermore, it's worth noting that nearly half of sponsors already incorporate these type of funds in their existing offerings, implying that the potential welfare gains only come from the half the sponsors not offering those type of funds as part of their plan menu.

Investors' outcome improves if, instead of mandating low-cost index funds, the policy mandates the inclusion of low-cost TDF. In this case, investor surplus increases by 11% and the average plan expense decreases by 23%. In magnitudes, the surplus for an investor with a $35,000 account balance increases by nearly $100 per year, an increase five times larger than the one obtained by mandating a low-cost S&P 500 tracker. Similarly, an household with a $70,000 income who contributes 10% to its 401(k) account would be saving $28,832 in fees over a 40 years period, an amount twice larger than the savings under a policy mandating the inclusion of low-cost S&P 500 trackers.

Why is it more effective to mandate the inclusion of low-cost TDFs than simply focusing on low-cost index funds? The primary reason lies in the benefit distribution: low-cost TDFs serve both active and, especially, inactive investors because they are used as qualified default

options. In contrast, mandating low-cost index funds predominantly benefits active investors, leaving inactive ones with no benefit in terms of reduced fees. Additionally, while TDFs often carry higher fees compared to index funds, the most affordable TDFs have expense ratios closely aligned with the ones charged by low-cost index funds. Take, for instance, the Fidelity Freedom Index series and the Vanguard Target Retirement series both offer TDFs with expense ratios under 10 basis points.

2.8.3 Capping funds' expenses

The last counterfactual I consider studies the effect of a 50 basis point expense ratio cap. Under this policy sponsors are allowed to include in their menu only funds with an expense ratio below 50 basis points. As a consequence, all funds whose marginal cost is higher than 50 basis point will exit the market. The latter are in most cases active funds, as more than 3/4 of passive funds and TDFs have an expense ratio below 50 basis points (Table 2.4).

This policy increases investor surplus by 14% and decreases the average plan expenses by 30%, corresponding to an increase in surplus of about 33 basis points and a decrease in plan expenses of about 15 basis points. A plan investor with a balance account of $35,000 enjoys an increase in surplus of about $120 dollars per-year whereas an investor contributing 10% of its $70,000 income is expected to save about $36,000 in fees over 40 years.

Perhaps not surprisingly, this policy is the most effective among the ones I considered thus far. On the one hand, it benefits inactive investors by eliminating the right-tail of expensive TDFs. On the other, it benefits active investors by ensuring that all investment options available are not excessively expensive.

Before concluding, one remark is in order. My analysis so far abstracted away from extensive margin considerations and implicitly assumed that sponsors would be always willing to provide a retirement plan to their employees. In practice, plan provision could be affected by these type of policies. For example, under a 50 basis point expense ratio cap,

it is likely that recordkeepers would lose revenues from revenue-sharing fees unless sponsors themselves compensate such loss by increasing their direct payments to their recordkeepers (Bhattacharya and Illannes (2022)). Some sponsors may be unwilling or might not have the resources to bear such costs and, consequently, might decide not to offer a retirement plan to their workers in the first place.

A plausible solution to minimize the extensive margin repercussions of these type of policies would be implementing such policies while at the same time subsidizing plan sponsors to incentivize plan provision. In practice, these type of subsidies have been already introduced in the 2019 SECURE Act to push small business to offer a retirement plan to their employees.

2.9 Conclusions

This paper proposes an equilibrium model of retirement plan menu choice, portfolio choice and fee competition between investment providers to uncover the factors contributing to the design high-cost employer-sponsored retirement plans and quantify the welfare effects of policies regulating plan design.

The model features a two-layer demand system where, in the first layer, sponsors choose their retirement plan and, in the second layer, plan investors form their retirement portfolio from the options available in their menu. On the supply side, investment funds compete by setting fees simultaneously while accounting for the two layers of demand. Funds compete for being included in a plan menu and for the plan assets.

I estimate the model using comprehensive data on retirement plan menus. Model estimates suggest that plan sponsors are less responsive to funds' fees than plan investors and value other fund characteristics, such as funds' affiliation with the plan recordkeeper. I use the estimated demand parameters to recover funds' price-cost margins and marginal costs from the implied Nash equilibrium conditions. Funds enjoy significant pricing power. This

is particularly evident for Target-Date-Funds (TDFs), who, although almost as cost-efficient as index funds, charge double the margins, with an implied median markup of about 34%.

In the last part of the paper, I consider four policy counterfactuals that regulate the design of plan menus and quantify their effect on plan investors' welfare. The first counterfactual shuts down sponsors' preferences for funds' affiliation. The second set of counterfactuals considers mandating the inclusion of low-cost options. The last counterfactual instead imposes a 50 basis point cap on funds' expense ratios. Among those, requiring the inclusion of low-cost TDFs and capping expense ratios are the most effective in improving investors' outcomes. Specifically, mandating the inclusion of low-cost TDFs increases investors' surplus by 11%, whereas a 50 basis points expense ratio cap leads to a 14% increase. Both policies also significantly reduce average plan expenses by 23% and 30%, respectively.

An important caveat of my analysis is that it abstracts from extensive margin considerations. In particular, it assumes that these types of regulations do not affect sponsors' incentives to offer a retirement plan in the first place. In practice, imposing expense ratio caps might reduce plan provision (Bhattacharya and Illannes (2022)). A practical solution would be pairing these policies with plan provision subsidies. Quantifying the optimal subsidy scheme and the effect of this combination of policies on plan investors' welfare is an important direction for future research.

2.10 Appendix: Additional Figures and Tables

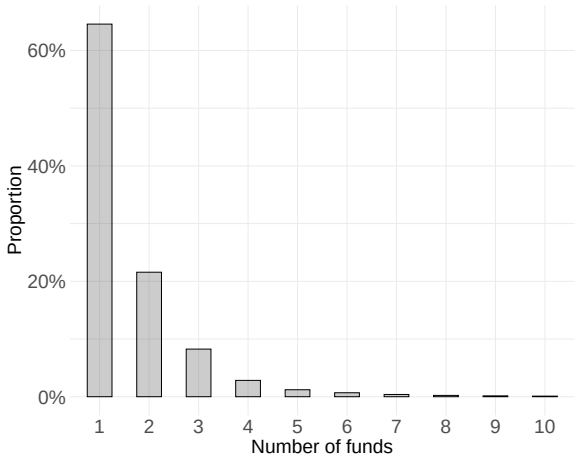

Figure 2.9: Distribution of number of options offered within investment category.

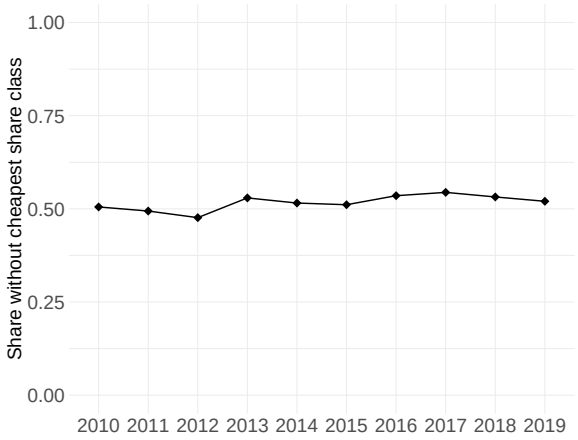

Figure 2.10: Within in fund × year share of employers who meet minimum investment required for cheapest share class but offer a more expensive one. The black line is the average share of employers without cheapest share class.

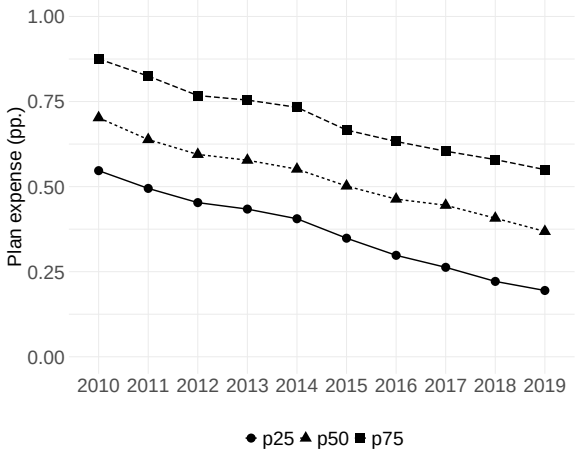

Figure 2.11: Distribution of average asset-weighted plan expense over time. Expenses are measured in percentage points

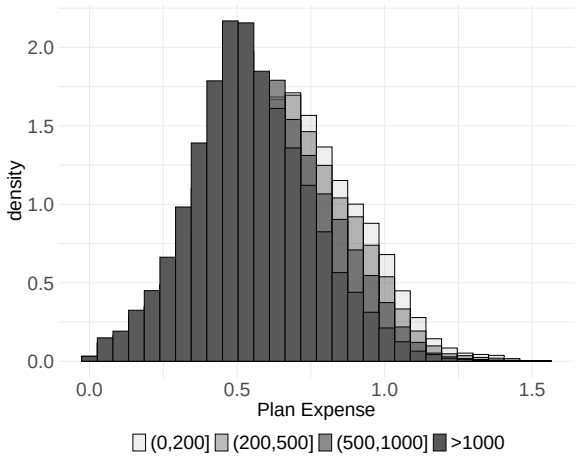

Figure 2.12: Distribution of plan expenses by plan size groups. Plan size is measured in number of participants.

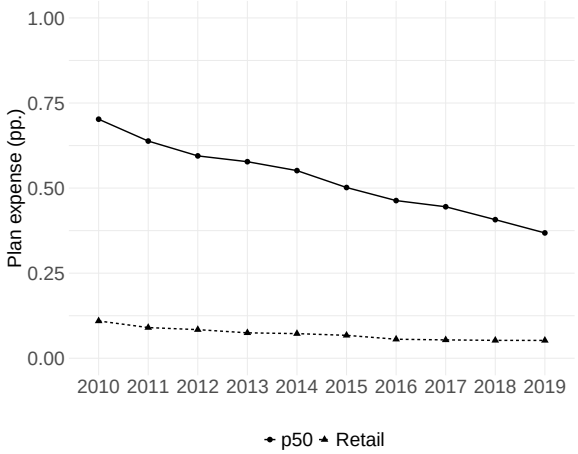

Figure 2.13: Median asset-weighted plan expense (dot-solid). Average expense ratio for a portfolio of Vanguard retail index funds (triangle-dashed). Expenses are measured in percentage points.

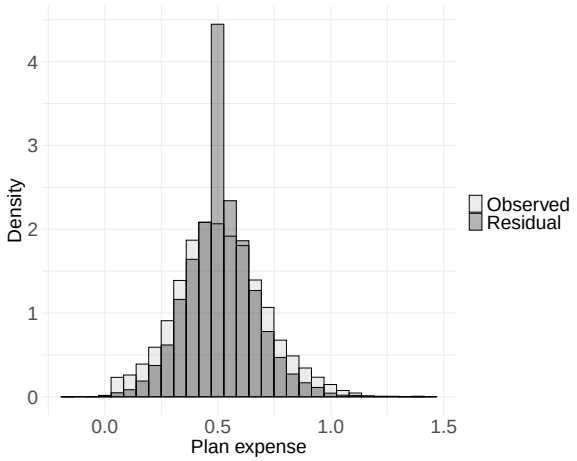

Figure 2.14: Within recordkeeper × six-digit NAICS dispersion in expenses. Other controls include plan assets, number of participants, and number of options.

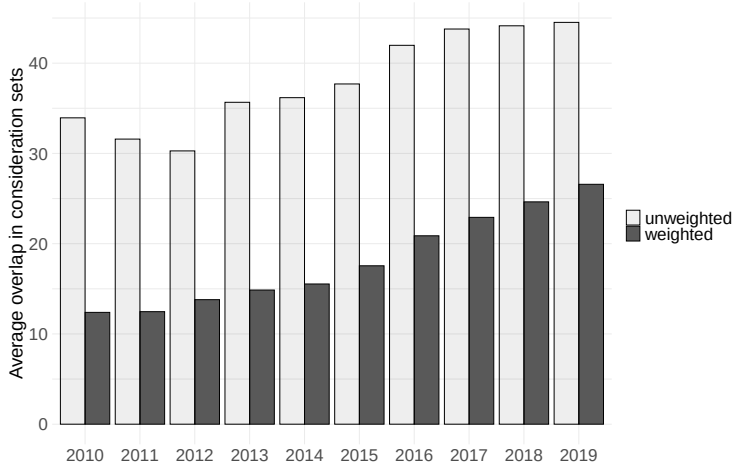

Figure 2.15: Average overlap in recordkeepers' network of funds. A fund belongs to a recordkeeper's network if it is offered in a plan managed by that same recordkeeper. The red bars represent the average fraction of funds that belong to the network of any two of the 10 largest recordkeepers. The turquoise bars represent the asset-weighted overlap.

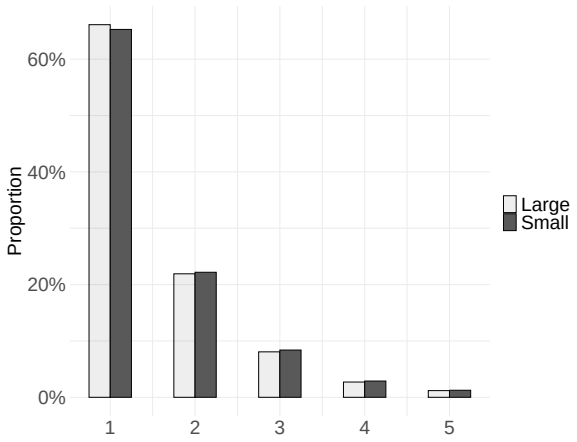

Figure 2.16: Distribution of number of options within category by plan size.

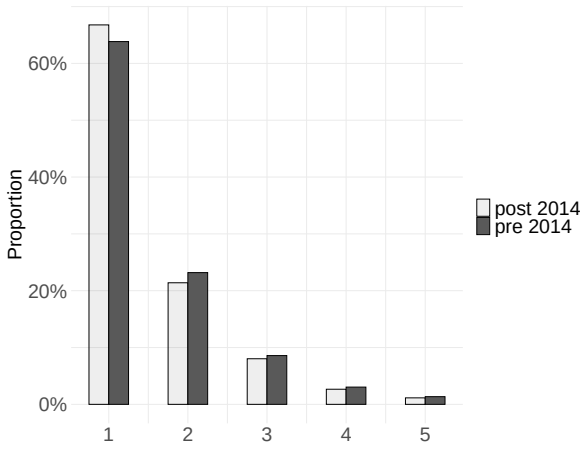

Figure 2.17: Distribution of number of options within category pre and post 2014.

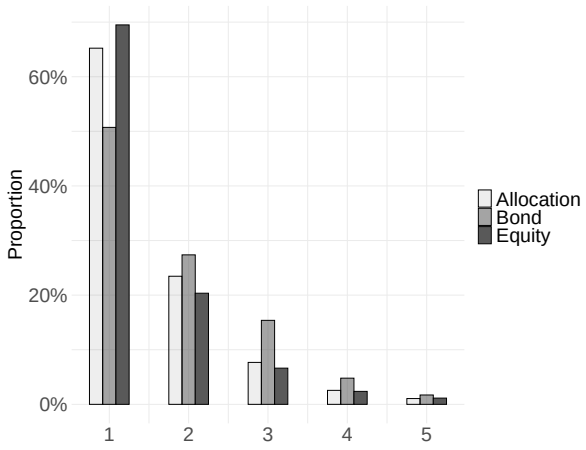

Figure 2.18: Distribution of number of options within category by asset class.

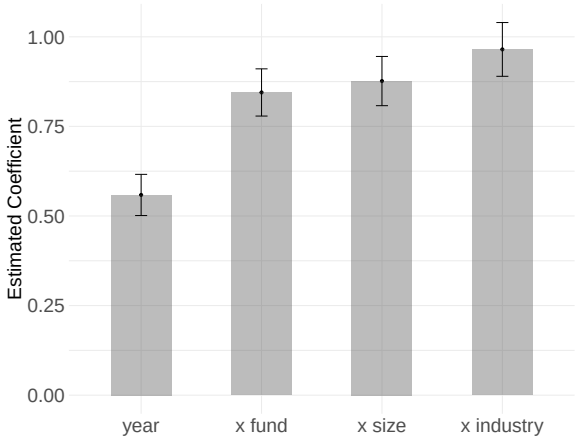

Figure 2.19: Coefficient from regressing log(inclusion probability) on affiliation dummy. Inclusion probability is the share of 401(k) plans offering a given fund. Inclusion probabilities are computed at the (year × size group × industry × recordkeeper) level for each fund. Size groups are based on the number of plan participants. Industry is the 2-digit NAICS.

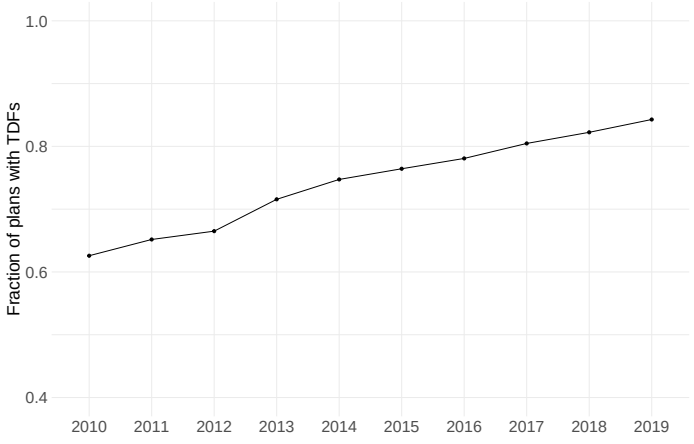

Figure 2.20: Share of retirement plans that offer at least one Target-Date-Fund (TDF).

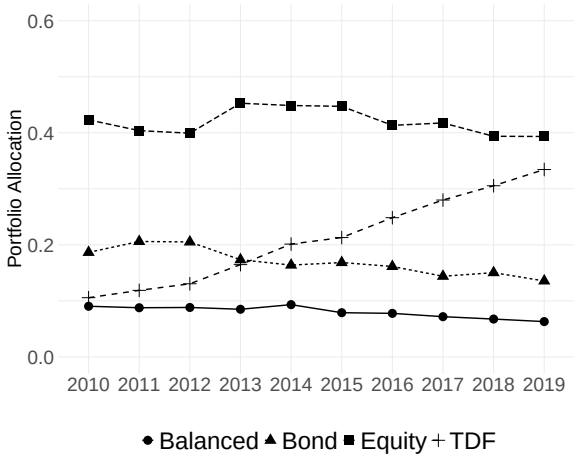

Figure 2.21: Average portfolio share across plan menus by asset class. Equity includes both US and International Equity funds. Balanced includes aggressive, moderate and conservative allocation funds that are not Target Date Funds (TDFs).

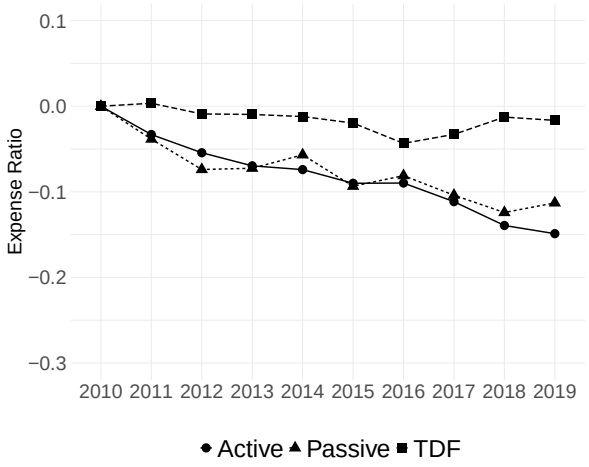

Figure 2.22: Secular decline in fees by fund type. The sample includes only funds available since 2010. The series for each type of fund has been shifted by the average fee as of 2010.

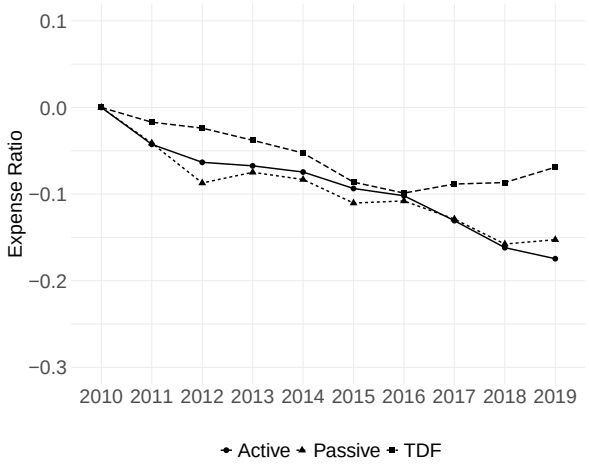

Figure 2.23: Secular decline in fees by fund type. The sample also includes funds introduced after 2010. The series for each type of fund has been shifted by the average fee as of 2010.

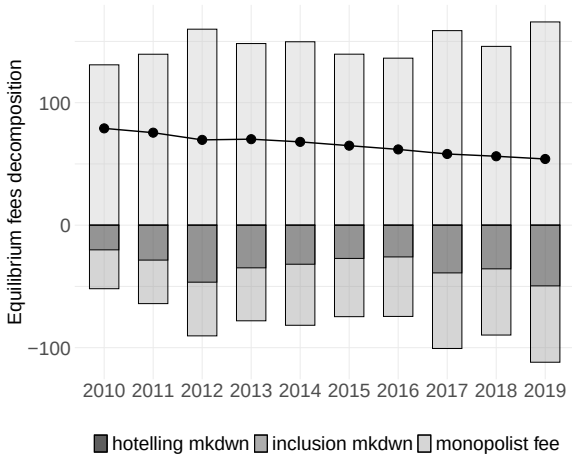

Figure 2.24: Cross-sectional decomposition of the average expense ratio into monopolist fee, hotelling markdown and plan inclusion markdown as defined in equation (2.20). Magnitudes are in basis points.

	Plan performance	Plan performance	Plan performance	Plan performance
Plan expenses	-0.76	-0.38	-0.44	-0.06
	(0.01)	(0.01)	(0.01)	(0.01)
Weighted	Y	Y	N	N
Year FE	N	Y	N	Y
R2	0.02	0.19	0.01	0.10

Table 2.5: Fund performance is the difference between fund return and the average category return. Plan performance is the average performance (possibly asset weighted) of all funds in the plan. Returns are gross of fees. Returns and fees are in percentage points.

	Exp. Ratio	Exp. Ratio	Exp. Ratio
log(# opt. in cat)	-0.047	-0.077	-0.051
	(0.001)	(0.001)	(0.002)
R2	0.07	0.27	0.62
Year FE	Y	Y	Y
Sponsor FE	N	Y	Y
Fund brand FE	N	N	Y

Table 2.6: Dependent variable is funds' expense ratio. Independent variable is the number of funds within an investment category. Expense ratios are in percentage points.

Year	Audited Plan Count	Potential Plan Count	Plan Coverage (%)	Audited Assets (bln)	Potential Assets (bln)	Asset Coverage (%)
2010	54626	66708	81.9	2,940	3,064	96.0
2011	52504	67394	77.9	3,023	3,134	96.4
2012	42372	67085	63.2	3,300	3,519	93.8
2013	42741	67514	64.8	4,086	4,158	98.3
2014	39360	68142	57.8	4,105	4,439	92.5
2015	40632	69038	58.9	4,225	4,437	95.2
2016	63464	70577	89.9	4,558	4,799	95.0
2017	68600	71855	95.5	5,504	5,553	99.1
2018	69424	73273	94.7	5,243	5,333	98.3
2019	71341	75057	95.0	6,289	6,372	98.7

Table 2.7: BrightScope Beacon data coverage. Assets are in billions.

Variable	N	Mean	SD	p5	p25	p50	p75	p95
Total Assets (mln.)	5615	190.736	1062.082	0.034	0.77	6.746	46.997	683.457
Portfolio share (avg.)	5615	2.896	3.821	0.119	0.887	1.795	3.279	10.039
Portfolio share (sd.)	5009	3.016	3.201	0.215	1.151	2.089	3.559	10.284
Fund-Plan turnover	5615	48.265	19.1	20.073	34.777	46.875	61.334	82.344
N. of share classes	5615	2.644	2.027	1	1	2	4	7

Table 2.8: Fund level summary statistics for the years 2010 to 2019. Each variable is first averaged (or summed in the case of 'total assets') within fund-year across plans, then within plan across years and tabulated across funds. The variable 'N' is the number of funds, excluding cash accounts and company stocks. Portfolio share (sd.) is the within fund-year standard deviation of the fund portfolio share across plans, which is then averaged within fund across years.

Employers preference parameters				
	Homogeneous preferences		Heterogeneous q	
	Mean	Marg. Effect (pp.)	Mean	Marg. Effect (pp.)
Expense Ratio (bp.)	-0.021	-0.008	-0.023	-0.009
	(0.002)	-	(0.002)	-
Affiliated (dummy)	0.823	0.328	0.852	0.330
	(0.044)	-	(0.044)	-
Target (dummy)	-0.414	-0.165	-0.463	-0.179
	(0.084)	-	(0.084)	-
Target × Affiliated	0.242	0.097	0.403	0.156
	(0.104)	-	(0.101)	-
Gross returns (pp.)	0.004	0.002	0.004	0.002
	(0.001)	-	(0.001)	-
Median fee elasticity	-1.83		-2.03	
q (Calibrated)	0.70		0.66	
GMM objective (df)	4.57 (2)		4.33 (2)	

Table 2.9: Two-step GMM estimates of plan sponsor preferences. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. For the heterogeneous q specification, q varies at the year-recordkeeper-category level.

Employers preference parameters								
	$(0, 200]$		$(200, 500]$		$(500, 1000]$		> 1000	
	Mean	Marg. Effect (pp.)	Mean	Marg. Effect (pp.)	Mean	Marg. Effect (pp.)	Mean	Marg. Effect (pp.)
Expense Ratio (bp.)	-0.019	-0.011	-0.017	-0.011	-0.024	-0.016	-0.026	-0.018
	(0.004)	-	(0.003)	-	(0.005)	-	(0.005)	-
Affiliated (dummy)	1.173	0.678	0.903	0.549	0.914	0.615	0.732	0.497
	(0.074)	-	(0.055)	-	(0.075)	-	(0.075)	-
Target (dummy)	-0.228	-0.132	-0.249	-0.151	-0.340	-0.229	-0.451	-0.307
	(0.133)	-	(0.105)	-	(0.160)	-	(0.176)	-
Target × Affiliated	-0.096	-0.055	0.147	0.090	-0.111	-0.075	-0.217	-0.148
	(0.142)	-	(0.122)	-	(0.161)	-	(0.170)	-
Gross returns (pp.)	0.002	0.001	0.004	0.003	0.005	0.003	0.003	0.002
	(0.002)	-	(0.002)	-	(0.002)	-	(0.003)	-
Median fee elasticity	-1.65		-1.50		-1.98		-2.08	
q (Calibrated)	0.70		0.70		0.70		0.70	
GMM objective (df)	3.91 (2)		2.91 (2)		1.56 (2)		2.76 (2)	

Table 2.10: Two-step GMM estimates of plan sponsor preferences for plans with number of participants below the median (small) and above the median (large). Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points.

Employers preference parameters				
	Before 2014		After 2014	
	Mean	Marg. Effect (pp.)	Mean	Marg. Effect (pp.)
Expense Ratio (bp.)	-0.014	-0.007	-0.029	-0.010
	(0.003)	-	(0.004)	-
Affiliated (dummy)	0.788	0.383	0.875	0.287
	(0.064)	-	(0.062)	-
Target (dummy)	-0.399	-0.194	-0.477	-0.157
	(0.126)	-	(0.117)	-
Target × Affiliated	0.381	0.185	0.158	0.052
	(0.156)	-	(0.142)	-
Gross returns (pp.)	0.003	0.001	0.004	0.001
	(0.002)	-	(0.002)	-
Median fee elasticity	-1.37		-2.40	
q (Calibrated)	0.70		0.70	
GMM objective (df)	4.88 (2)		6.38 (2)	

Table 2.11: Two-step GMM estimates of plan sponsor preferences for the pre 2014 and post 2014 subsamples. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points.

Sponsors preference parameters				
	No inertia		Yes inertia	
	Mean	Marg. Effect (pp.)	Mean	Marg. Effect (pp.)
Expense Ratio (bp.)	-0.023	-0.009	-0.015	-0.006
	(0.003)	-	(0.003)	-
Affiliated (dummy)	0.828	0.338	0.510	0.208
	(0.045)	-	(0.045)	-
Target (dummy)	-0.453	-0.185	-0.097	-0.040
	(0.090)	-	(0.088)	-
Target × Affiliated	0.236	0.096	0.265	0.108
	(0.101)	-	(0.106)	-
Gross returns (pp.)	0.003	0.001	0.004	0.002
	(0.002)	-	(0.002)	-
log(Lag inclusion prob.)	-	-	0.259	0.106
	-	-	(0.003)	-
Median fee elasticity	-1.97		-1.30	
q (Calibrated)	0.70		0.70	
GMM objective (df)	6.68 (2)		3.41 (2)	

Table 2.12: Two-step GMM estimates of plan sponsor preferences accounting for inertia in menu choices. Robust standard errors are reported in parentheses. Year, category, passive and fund brand fixed effects are included. For the marginal effects, inclusion probabilities are in percentage points. Sample is restricted to plans observed for at least two consecutive years.

	Projected R2	R2
alpha	0.006	0.051
beta	0.039	0.083
category	0.143	0.182
alpha | category	0.000	0.182
beta | category	0.002	0.184
category | fund provider	0.117	0.203

Table 2.13: Dependent variable is plan-level portfolio allocations. All specification include plan × year fixed effects. Beta are 3 Fama-French plus Momentum and 3 bond factors.

Plan investors preference parameters						
	OLS			IV-Turnover	IV-Hausmann	ME
Expense ratio ($\tilde{\gamma}$)	-0.004	-0.011	-0.012	-0.044	-0.040	-0.098
	(0.001)	(0.001)	(0.001)	(0.004)	(0.003)	
Affiliated ($\tilde{\beta}_1$)	0.001	0.004	0.017	0.015	0.016	0.034
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	
Gross returns ($\tilde{\beta}_2$)	0.002	0.012	0.013	0.020	0.020	0.043
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	
Fraction inactive (δ)	0.267	0.290	0.407	0.371	0.375	-
	(0.006)	(0.006)	(0.008)	(0.009)	(0.008)	
				1st Stage	1st stage	
Turnover ratio	-	-	-	0.028	-	-
				(0.000)		
Hausman IV	-	-	-	-	-0.208	-
					(0.002)	
Median fee elasticity	-0.225	-0.645	-0.688	-2.517	-2.263	-
Median fee elasticity (active investors)	-0.306	-0.909	-1.162	-4.002	-3.621	-
Fstat	-	-	-	82.55	86.69	-
R2	0.24	0.25	0.42	0.82	0.82	-
Fund brand FE	N	Y	Y	Y	Y	-
Employer FE	N	N	Y	Y	Y	-

Table 2.14: Estimates of plan investors preferences. All specifications include year, category and passive fixed effects. Expense ratios are in percentage points (pp.). R2 for IV columns is first stage. ME are the (median) marginal effects for portfolio allocations in pp. for a basis point increase in expenses or a pp. increase gross returns. Turnover and Hausman IV are standardized.

Equilibrium decomposition of observed fees			
Fee	Monopolist fee	Hotelling markdown	Plan inclusion markdown
65.75	119.44	25.15	28.54

Table 2.15: Decomposition of fees following equation (2.20). All magnitudes are in basis points. The figures shown are averages across time and funds.

	Investor Surplus (bp.)	Average plan expense (bp.)	Fee savings/year ($)	Fee savings/40 years ($)
Status Quo	238	51	-	-
No Affiliation Preference	235	50	-	-
Low-cost Index Fund	243	46	18	11,897
Low-cost TDF	265	39	42	28,832
Expense cap (50 bp.)	271	36	53	36,192

Table 2.16: Investor surplus and average plan expense under different counterfactual policies. Magnitudes are in basis points. Savings are relative to the status quo. Fee savings per-year assumes a retirement account balance of $35,000. Fee savings over 40 years assumes an annual income of $70,000, contribution rate of 10% and an annual return of 6%. Expense ratio cap is at 60 basis points.

2.11 Appendix: Model derivations

In this Appendix I provide more formal derivations of the results introduced in the main text.

Derivation of ranking probabilities. Consider the simple example in the main text where we have four options $\{j, k, l, m\}$ and we want to compute the probability that option j is ranked 2nd in terms of sponsors' indirect utilities. For simplicity I drop sponsor subscript p. The probability that j is ranked 2nd equals to the sum of all possible utility rankings in which u_j is the second highest:

$$\begin{aligned}\phi_j^2 = &\Pr\{u_k > u_j > u_l > u_m\} + \Pr\{u_k > u_j > u_m > u_l\} \\ &+ \Pr\{u_l > u_j > u_k > u_m\} + \Pr\{u_l > u_j > u_m > u_k\} \\ &+ \Pr\{u_m > u_j > u_k > u_l\} + \Pr\{u_m > u_j > u_l > u_k\}. \end{aligned} \tag{2.31}$$

Next, consider any of the six rankings above, say the first one and note that

$$\Pr\{u_k > u_j > u_l > u_m\} = \frac{\exp(V_k)}{\sum_{s\in\{j,k,l,m\}} \exp(V_s)} \cdot \frac{\exp(V_j)}{\sum_{s'\in\{j,l,m\}} \exp(V_{s'})} \cdot \frac{\exp(V_l)}{\sum_{s''\in\{l,m\}} \exp(V_{s''})} \tag{2.32}$$

Expression (2.32) can be derived analytically by integrating over the T1EV extreme value shocks and is known as ranked-ordered-logit (ROL),[52] which can be interpreted as a sequential multinomial logit decision problem.

Expression (2.32) applies analogously to all six terms in (2.31) and implies that ϕ_j^2 only depends on how the first two choices are ranked but not on the order of the 3rd and 4th choices. To see this consider the sum of the first two terms on the RHS of (2.31) and note

52. For more details see Beggs, Cardell and Hausman (1981).

that from (2.32) we can factor out the first two factors of each addend and that the sum of the last factors equals. Overall we obtain

$$\begin{aligned}&\Pr\{u_k > u_j > u_l > u_m\} + \Pr\{u_k > u_j > u_m > u_l\} = \\ &= \frac{\exp(V_k)}{\sum_{s\in\{j,k,l,m\}} \exp(V_s)} \cdot \frac{\exp(V_j)}{\sum_{s'\in\{j,l,m\}} \exp(V_{s'})}.\end{aligned}$$

Applying the same steps for all three lines in (2.31), the three term expression in the main text obtains.

Derivation of unconditional plan inclusion probability. Letting λ_{gp} the probability that p chooses to offer category g, $q(1-q)^{n-1}$ the probability that p includes n funds from category g and $\phi_j^{1:n}$ the probability that j is chosen conditional on n options and g being offered, the unconditional probability that j ends up being in p's retirement plan can be rearranged as follows

$$\begin{aligned}\phi_{jp} &= \lambda_{gp} \cdot \left(\sum_{n=1}^{\infty} q(1-q)^{n-1} \phi_{jp}^{1:n} \right) \\ &= \lambda_{gp} \cdot \left(\sum_{n=1}^{\infty} q(1-q)^{n-1} \sum_{z=1}^{n} \phi_{jp}^{z} \right) \\ &= \lambda_{gp} \cdot \sum_{z=1}^{\infty} \sum_{n=z}^{\infty} q(1-q)^{n-1} \phi_{jp}^{z} \\ &= \lambda_{gp} \cdot \sum_{z=1}^{\infty} \phi_{jp}^{z} (1-q)^{z-1} \sum_{n=z}^{\infty} q(1-q)^{n-z} \\ &= \lambda_{gp} \cdot \sum_{z=1}^{\infty} \phi_{jp}^{z} (1-q)^{z-1}\end{aligned}$$

where the latter coincides with expression (2.6) provided in the main text.

Heterogeneous individual investors. Let us consider the case in which individual investors have heterogeneous preferences. Formally, let A_i, δ_i, $\boldsymbol{\beta}_i$ and γ_i be individual specific.

Moreover, assume that there is a subset $D \subset J_p$ of default funds and denote by d_i investor i's default fund. Under these assumptions i's demand system is given by

$$\boldsymbol{a}_i(\boldsymbol{f}) = \begin{cases} \boldsymbol{e}_{d_i} & \text{if } i \text{ defaults} \\ \frac{1}{\gamma_i}(I+G)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{f}) & \text{o.w} \end{cases} \tag{2.33}$$

where to save on notation I defined $\boldsymbol{\mu}_i \equiv W\boldsymbol{\beta}_i + \boldsymbol{\xi}_i$. To obtain the aggregate demand system let us define the following weighted averages

$$\begin{aligned} \boldsymbol{\mu}_p &\equiv \sum_{i \in I_p} \frac{(1-\delta_i)\gamma_i^{-1}}{\sum_{i \in I_p}(1-\delta_i)\gamma_i^{-1}} \boldsymbol{\mu}_i \\ \gamma_p &\equiv \left(\sum_{i \in I_p} \frac{A_i}{A_p} \frac{1-\delta_i}{\gamma_i} \right)^{-1} \\ \delta_{dp} &\equiv \left(\frac{\sum_{i \in I_{dp}} A_i}{A_p} \right) \sum_{i \in I_{dp}} \frac{A_i}{\sum_{i \in I_{dp}} A_i} \delta_i \end{aligned}$$

where I_p is the set of plan p investors and I_{dp} is the set of investors that have fund d as default option. Then taking the horizontal sum of the $\boldsymbol{a}_i$ it is easy to check that the plan level demand system is given by

$$\boldsymbol{s}_p(\boldsymbol{f}; \boldsymbol{\eta}_p) = \sum_{d \in D} \delta_{dp} \mathbf{e}_d + \frac{1}{\gamma_p}(I + G_p)^{-1}(\boldsymbol{\mu}_p - \boldsymbol{f}).$$

Overall, the estimated parameters from the aggregate demand system $\boldsymbol{\eta}_p = (\delta_{dp}, \gamma_p, \boldsymbol{\mu}_p)$ are weighted averages of the underlying individual investors parameters.

Derivation of aggregate demand system in (2.12). To show that equations (2.10) and

(2.12) are equivalent it is enough to note that

$$(I - \tilde{G}(I + \tilde{G}'\tilde{G})^{-1}\tilde{G}')(I + G) = \tag{2.34}$$

$$= (I - \tilde{G}(I + \tilde{G}'\tilde{G})^{-1}\tilde{G}')(I + \tilde{G}\tilde{G}') = \tag{2.35}$$

$$= I - \tilde{G}(I + \tilde{G}'\tilde{G})^{-1}\tilde{G}' + \tilde{G}\tilde{G}' - \tilde{G}(I + \tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{G}\tilde{G}' \tag{2.36}$$

$$= I + \tilde{G}(I - (I + \tilde{G}'\tilde{G})^{-1} - (I + \tilde{G}'\tilde{G})^{-1}\tilde{G}'\tilde{G})\tilde{G}' \tag{2.37}$$

$$= I + \tilde{G}(I + \tilde{G}'\tilde{G})^{-1}(I + \tilde{G}'\tilde{G} - I - \tilde{G}'\tilde{G})\tilde{G}' = I \tag{2.38}$$

Next, I show that $\kappa_{jj} \in (0, 1)$ and $\kappa_{jl} \in (-1, 1)$ for all j and l. To see this, consider note that by construction $\mathcal{K}_x$ and $I - \mathcal{K}_x$ are positive definite matrices. The let e_j be the jth unit vector and note the definition of positive definite matrix implies that

$$\kappa_{jj} = e_j'\mathcal{K}_x e_j > 0 \quad \text{and} \quad 1 - \kappa_{jj} = e_j'(I - \mathcal{K}_x)e_j > 0. \tag{2.39}$$

Next, take any (j, l) pair with $j \neq l$ and note that

$$\kappa_{jj} + \kappa_{ll} - 2\kappa_{jl} = (e_j - e_l)'\mathcal{K}_x(e_j - e_l) > 0 \tag{2.40}$$

where the first equality exploits the fact that $\mathcal{K}_x$ is symmetric. From (2.40) and the fact that $\kappa_{jj} < 1$ for all j, we can conclude that $\kappa_{jl} < 1$. To show that $\kappa_{jl} > -1$ it is enough to repeat the previous argument using $\kappa_{jj} + \kappa_{ll} + 2\kappa_{jl}$.

Derivation of expected κ_{jl} under biased beliefs about q. Suppose that fund j believes that sponsors will include at most one fund per investment category. This means that funds evaluate inclusion probabilities assuming that $q = 1$ and will assign positive probabilities only to menus S that include at most one fund per category.

Denoting by G_S the set of categories included in menu S, by j_g a generic option from category g and by g_j the investment category fund j belongs to, the probability that such

menu $S \in \mathcal{S}_{jp}$ is chosen by sponsor p can be factored us

$$\phi_p(S) = \phi_{jp} \prod_{g \in G_S/g_j} \phi_{j_g p} \prod_{j_{g'} p} (1 - \lambda_{j_{g'} p}) \tag{2.41}$$

where the factorization is a consequence of the fact that inclusion decision are made independently across investment categories.

Next, consider fund j and fund l and note that

$$\bar{\kappa}_{jl} = \sum_{S \in \mathcal{S}_{jp}} \frac{\phi_p(S)}{\phi_{jp}} \kappa_{jl}^S \tag{2.42}$$

$$= \sum_{S \in \mathcal{S}_{jp}} \left(\prod_{g \in G_S/g_j} \phi_{j_g p} \prod_{j_{g'} p} (1 - \lambda_{j_{g'} p}) \right) \kappa_{jl}^S \tag{2.43}$$

which does not depend on f_j because inclusion probabilities of competitors funds belonging to different categories do not depend on f_j.

Derivation of equilibrium fees decomposition. Recall the system of Bertrand FOCs' derived in the main text

$$\bar{\delta} \boldsymbol{e}_d + (I - \tilde{\mathcal{K}}) (\tilde{\boldsymbol{\mu}} - \boldsymbol{f}) - \boldsymbol{\iota} - (I - diag(\tilde{\mathcal{K}}))(\boldsymbol{f} - \boldsymbol{c}) = 0 \tag{2.44}$$

which can be rearranged as

$$2 \left(I - \frac{diag(\tilde{\mathcal{K}})}{2} - \frac{\tilde{\mathcal{K}}}{2} \right) \boldsymbol{f} = (I - \tilde{\mathcal{K}}) \tilde{\boldsymbol{\mu}} + (I - diag(\tilde{\mathcal{K}})) \boldsymbol{c} - \boldsymbol{\iota}. \tag{2.45}$$

Next define

$$\tilde{\boldsymbol{\iota}} \equiv \left(I - \frac{diag(\tilde{\mathcal{K}})}{2} - \frac{\tilde{\mathcal{K}}}{2} \right)^{-1} \frac{\boldsymbol{\iota}}{2} \tag{2.46}$$

and rewrite the system of FOCs as

$$\begin{aligned}
\boldsymbol{f} &= \frac{1}{2}\left(I - \frac{diag(\tilde{\mathcal{K}})}{2} - \frac{\tilde{\mathcal{K}}}{2}\right)^{-1}\left[(I-\tilde{\mathcal{K}})\tilde{\boldsymbol{\mu}} + (I - diag(\tilde{\mathcal{K}}))\boldsymbol{c}\right] - \tilde{\boldsymbol{\iota}} \\
&= \boldsymbol{c} + \frac{1}{2}\left(I - \frac{diag(\tilde{\mathcal{K}})}{2} - \frac{\tilde{\mathcal{K}}}{2}\right)^{-1}(I-\tilde{\mathcal{K}})(\tilde{\boldsymbol{\mu}} - \boldsymbol{c}) - \tilde{\boldsymbol{\iota}} \\
&= \boldsymbol{c} + \frac{1}{2}\left(I - diag(\tilde{\mathcal{K}}) - \frac{\tilde{\mathcal{K}} - diag(\tilde{\mathcal{K}})}{2}\right)^{-1}(I-\tilde{\mathcal{K}})(\tilde{\boldsymbol{\mu}} - \boldsymbol{c}) - \tilde{\boldsymbol{\iota}} \\
&= \frac{\tilde{\boldsymbol{\mu}} + \boldsymbol{c}}{2} - \left(I - diag(\tilde{\mathcal{K}}) - \frac{\tilde{\mathcal{K}} - diag(\tilde{\mathcal{K}})}{2}\right)^{-1}\frac{\tilde{\mathcal{K}} - diag(\tilde{\mathcal{K}})}{2}\frac{\tilde{\boldsymbol{\mu}} - \boldsymbol{c}}{2} - \tilde{\boldsymbol{\iota}} \\
&= \frac{\tilde{\boldsymbol{\mu}} + \boldsymbol{c}}{2} - (I - diag(\tilde{\mathcal{K}}))^{-1/2}\left(I - \frac{G(\tilde{\mathcal{K}})}{2}\right)^{-1}\frac{G(\tilde{\mathcal{K}})}{2}(I - diag(\tilde{\mathcal{K}}))^{1/2}\frac{\tilde{\boldsymbol{\mu}} - \boldsymbol{c}}{2} - \tilde{\boldsymbol{\iota}}
\end{aligned}$$

where

$$G(\tilde{\mathcal{K}}) \equiv (I - diag(\tilde{\mathcal{K}}))^{-1/2}(\tilde{\mathcal{K}} - diag(\tilde{\mathcal{K}}))(I - diag(\tilde{\mathcal{K}}))^{-1/2}. \tag{2.47}$$

Expression (2.21) in the main text obtains by setting $diag(\tilde{\mathcal{K}}) = k_0 I$. For the case in which there is a default fund the same steps apply after redefining $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}} + \bar{\delta}(I - \tilde{\mathcal{K}})^{-1}\boldsymbol{e}_d$.

Estimation algorithm. Before starting the estimation, I draw a vector $\boldsymbol{\nu}_s$ of random taste parameters for $s = 1, ..., S$ simulated sponsors from a normal $N(\mathbf{0}, I)$ and store it. Then the algorithm proceeds as follows.

Step 0: Guess Γ_θ

Step 1: For a given guess of the vector of the mean utility mean utility $\bar{\boldsymbol{v}}^{(k)}$. Compute the following variables for each fund j, market t

1.1 for each simulated sponsor s calculate the following objects

- the probability that j's category g is included by sponsor s

$$\lambda_{gst}(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) = \frac{\sum_{l \in g} \exp(\bar{v}_{lt}^{(k)} + \boldsymbol{w}_{lt}' \Gamma_\theta \boldsymbol{\nu}_s)}{1 + \sum_{l \in g} \exp(\bar{v}_{lt}^{(k)} + \boldsymbol{w}_{lt}' \Gamma_\theta \boldsymbol{\nu}_s)}$$

- for $n = \{1, 2, 3\}$ calculate the ranking probabilities

$$\phi_{jst}^n(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) = \sum_{(j_1,\ldots,j_{n-1}) \in g/\{j\}} \prod_{n'=1}^{n-1} \frac{\exp(\bar{v}_{j_{n'}t}^{(k)} + \boldsymbol{w}_{j_{n'}t}' \Gamma_\theta \boldsymbol{\nu}_s)}{\sum_{n''=n'}^{N_{gp}} \exp(\bar{v}_{j_{n''}t}^{(k)} + \boldsymbol{w}_{j_{n''}t}' \Gamma_\theta \boldsymbol{\nu}_s)} \cdot \frac{\exp(\bar{v}_{jt}^{(k)} + \boldsymbol{w}_{jt}' \Gamma_\theta \boldsymbol{\nu}_s)}{\sum_{n'''=n}^{N_{gp}} \exp(\bar{v}_{j_{n'''}t}^{(k)} + \boldsymbol{w}_{j_{n'''}t}' \Gamma_\theta \boldsymbol{\nu}_s)}$$

- compute the probability that sponsor s includes fund j

$$\phi_{jst}^{1:3}(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) = \phi_{jst}^1(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) + (1 - q_t)\phi_{jst}^2(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) + (1 - q_t)^2 \phi_{jst}^3(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)})$$

where q_t is calibrated to match the empirical distribution of the number of options within investment category in market t. The typical values for q_t are of 0.7 or higher so that $(1 - q_t)^{n-1}$ decays quite fast in n. In simulations, I find that stopping at $n = 3$ works well in recovering the true parameters. Moreover the computational burden of calculating ϕ_{jst}^n for $n \geq 4$ is non negligible.

1.2 approximate the RHS of (2.22) with

$$\phi_{jt}^S(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) = \frac{1}{S} \sum_{s=1}^{S} \lambda_{gst}(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)}) \cdot \phi_{jst}^{1:3}(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)})$$

Step 2: For each t update the mean utility vector by computing $\bar{\boldsymbol{v}}_t^{(k+1)}$ as

$$\bar{\boldsymbol{v}}_t^{(k+1)} = \bar{\boldsymbol{v}}_t^{(k)} + \log(\hat{\boldsymbol{\phi}}_t(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)})) - \log(\boldsymbol{\phi}_t^S(\Gamma_\theta, \bar{\boldsymbol{v}}_t^{(k)})) \tag{2.48}$$

Step 3: For each t Repeat Step 1 and Step 2 until

$$||\bar{\boldsymbol{v}}_t^{(k+1)} - \bar{\boldsymbol{v}}_t^{(k)}||_\infty < \epsilon \tag{2.49}$$

for some tolerance level ϵ.

Step 4: Recover sponsors' preference parameters $\boldsymbol{\mu}_\theta$ that enter sponsors' utility linearly

$$\boldsymbol{\mu}_\theta(\Gamma_\theta) = (W'Z(Z'Z)^{-1}Z'W)^{-1}W'Z(Z'Z)^{-1}Z'\boldsymbol{v}(\Gamma_\theta) \tag{2.50}$$

where W is a matrix of funds' characteristics with number of rows equal to the total number of observations $\bar{N} = \sum_t N_t$ and number of columns equal to the number of characteristics and Z is including both excluded and included instruments.

Step 5: Recover demand residuals

$$\boldsymbol{\zeta}(\Gamma_\theta) = \boldsymbol{v}(\Gamma_\theta) - W\boldsymbol{\mu}_\theta \tag{2.51}$$

and compute the GMM norm

$$\boldsymbol{\zeta}(\Gamma_\theta)'Z\Omega(\Gamma_\theta)Z'\boldsymbol{\zeta}(\Gamma_\theta) \tag{2.52}$$

where $\Omega(\Gamma_\theta) = (Z'Z)^{-1}$ in the first GMM estimation step and then is updated to $\Omega(\Gamma_\theta) = Z'diag(\boldsymbol{\zeta}(\Gamma_\theta)^2)Z = \sum_{j,t}\zeta_{jt}^2(\Gamma_\theta)\boldsymbol{Z}_{jt}\boldsymbol{Z}_{jt}'$.

Interior equilibrium existence and uniqueness. Consider first the case in which funds know with certainty which plan menu will include them. Formally, this means that $\phi_{jp} = 1$ when p includes fund j and zero otherwise which further implies that the $\boldsymbol{\iota}$ in equation (2.21)

equals zero. Given this, the system of funds' best replies becomes linear in $\boldsymbol{f}$:

$$(I - \tilde{\mathcal{K}})\boldsymbol{f} + (I - diag(\tilde{\mathcal{K}}))\boldsymbol{f} = (I - \tilde{\mathcal{K}})\tilde{\boldsymbol{\mu}} + (I - diag(\tilde{\mathcal{K}}))\boldsymbol{c} \tag{2.53}$$

because inclusion probabilities do not depend on fees anymore and with that also $\mathcal{K}$, $\tilde{\boldsymbol{\mu}}$ do not depend on fees. A well-defined solution is then guaranteed to exist as long as $(I - \tilde{\mathcal{K}})$ is invertible. Moreover, linearity would also imply that such solution is unique.

What we do not know is whether this solution is such that equilibrium fees are non-negative. In what follows I show that the following dominance-diagonal condition

$$(1 - \tilde{\kappa}_{jj})(\tilde{\mu}_j - c_j) > \sum_{k \neq j} |\tilde{\kappa}_{jk}|(\tilde{\mu}_k - c_k) \quad \text{all } j \tag{2.54}$$

implies that the system of best replies is a self-map over the interior of the set $\times_{j\in\{1,...,J\}}[c_j, \tilde{\mu}_j]$ which ensures that equilibrium fees are positive and above marginal costs. I start by defining the following linear operator $T : \mathbb{R}^J \to \mathbb{R}^J$ whose j-th component is fund j's best reply

$$T_j(\boldsymbol{f}) \equiv \frac{1}{2}\left[\tilde{\mu}_j - \sum_{k \neq j} \frac{\tilde{\kappa}_{jk}}{1 - \tilde{\kappa}_{jj}}\mu_k + c_j + \sum_{k \neq j} \frac{\tilde{\kappa}_{jk}}{1 - \tilde{\kappa}_{jj}} f_k\right]. \tag{2.55}$$

Assumption (2.54) implies that T is a self-map in the interior of $\times_{j\in\{1,...,J\}}[c_j, \tilde{\mu}_j]$. To see this take any $\boldsymbol{f} \in \times_{j\in\{1,...,J\}}[c_j, \tilde{\mu}_j]$ and note that

$$c_j < T_j(\boldsymbol{f}) < \tilde{\mu}_j \tag{2.56}$$

$$\Leftrightarrow \left|\sum_{k \neq j} \frac{\tilde{\kappa}_{jk}}{1 - \tilde{\kappa}_{jj}} \frac{\tilde{\mu}_k - c_k}{\tilde{\mu}_j - c_j} \frac{\tilde{\mu}_k - f_k}{\tilde{\mu}_k - c_k}\right| < 1 \tag{2.57}$$

$$\Leftrightarrow \sum_{k \neq j} \frac{|\kappa_{jk}|}{1 - \tilde{\kappa}_{jj}} \frac{\tilde{\mu}_k - c_k}{\tilde{\mu}_j - c_j} \frac{\tilde{\mu}_k - f_k}{\tilde{\mu}_k - c_k} < 1 \tag{2.58}$$

is always satisfied when (2.54) holds and $\boldsymbol{f} \in \times_{j\in\{1,...,J\}}[c_j, \tilde{\mu}_j]$. Because T maps a closed

and bounded set into itself, Bower fixed point theorem implies that there exists an $\boldsymbol{f}^* \in \times_{j\in\{1,...,J_p\}}[c_j, \mu_j]$ such that $T(\boldsymbol{f}^*) = \boldsymbol{f}^*$. Moreover, from (2.58) we know that such fixed point is interior and unique, because T is linear.

We can also show that in this interior equilibrium each fund manages a positive amount of asset. To see this, consider fund j's first order condition evaluated at the optimum

$$s_j(\boldsymbol{f}^*) - (1 - \tilde{\kappa}_{jj})(f_j^* - c_j) = 0 \tag{2.59}$$

which implies that

$$s_j(\boldsymbol{f}^*) = (1 - \tilde{\kappa}_{jj})(f_j^* - c_j) > 0 \tag{2.60}$$

where the latter inequality holds because we just proved that $f_j^* > c_j$ for all j and $(1-\tilde{\kappa}_{jj}) > 0$ follows from the fact that $I - \mathcal{K}$ is positive definite and that $\tilde{\kappa}_{jj}$ is just an average of the κ_{jj} across plans:

$$\tilde{\kappa}_{jj} = \bar{\phi}_j^{-1} \int \mathbf{1}\{j \in S_p\}(1 - \delta_p)\gamma_p^{-1} A_p \kappa_{jj}^{S_p} dF_p$$

with

$$\bar{\phi}_j = \int \mathbf{1}\{j \in S_p\}(1 - \delta_p)\gamma_p^{-1} A_p dF_p. \tag{2.61}$$

Lastly I show that the equilibrium is stable. To see this define the following variable

$$\hat{f}_j \equiv \frac{f_j - c_j}{\tilde{\mu}_j - c_j} \tag{2.62}$$

and note that fund j best reply can be rewritten as

$$\hat{f}_j = \frac{1}{2}\left[1 - \sum_{k \neq j} \frac{\tilde{\kappa}_{jk}}{1 - \tilde{\kappa}_{jj}} \frac{\tilde{\mu}_k - c_k}{\tilde{\mu}_j - c_j}\left(1 - \frac{f_k - c_k}{\tilde{\mu}_k - c_k}\right)\right] \tag{2.63}$$

Defining the linear mapping on the above RHS as $\hat{T} : R^J \to R^J$, it can be shown that this mapping is a self-map into $[0,1]^J$ under assumption (2.54). Additionally, this mapping is a contraction in the L_∞ norm.

$$||\hat{T}(\hat{\boldsymbol{f}}_1) - \hat{T}(\hat{\boldsymbol{f}}_0)||_\infty = \max_j |\hat{T}_j(\hat{\boldsymbol{f}}_1) - \hat{T}_j(\hat{\boldsymbol{f}}_0)| \tag{2.64}$$

$$= \max_j \left| \sum_{k \neq j} \frac{\kappa_{jk}}{1 - \kappa_{jj}} \frac{\mu_k - c_k}{\tilde{\mu}_j - c_j} (\hat{f}_{1k} - \hat{f}_{0k}) \right| \tag{2.65}$$

$$\leq \max_j \sum_{k \neq j} \frac{|\kappa_{jk}|}{1 - \kappa_{jj}} \frac{\mu_k - c_k}{\mu_j - c_j} |\hat{f}_{1k} - \hat{f}_{0k}| \tag{2.66}$$

$$< \max_k |f_{1k} - f_{0k}| \tag{2.67}$$

which ensures that the unique equilibrium is also stable.

Next, I consider the case in which sponsor preferences are homogeneous $\theta_p \equiv \theta$ and funds do not know with certainty whether or not they will be included in sponsors' retirement menus (e.g., $\phi_j \in (0,1)$). For simplicity, I also assume that there is only one recordkeeper or equivalently that all recordkeerpers have the same network of funds. Under this assumptions, I will show that a Nash-Bertrand equilibrium exists when funds believe that sponsors include at most one fund per category and the previous dominance diagonal condition holds.

To start with, note that fund j problem in any given period simplifies to

$$\max_{f_j} \quad P \cdot (f_j - c_j) \cdot \phi_j(\boldsymbol{f}_j; \boldsymbol{\theta}) \int s_{jp}(\boldsymbol{f}; \boldsymbol{\eta}_p) A_p dF(\boldsymbol{\eta}_p, A_p) \tag{2.68}$$

where fund j expected portfolio share is given by

$$\begin{aligned} s_{jp}(\boldsymbol{f};\eta_p) &= \sum_{S\in\mathcal{S}_j} \tilde{\gamma}_p \frac{\phi(S;\boldsymbol{\theta})}{\phi_j(\boldsymbol{f};\boldsymbol{\theta})} \left[(1-\kappa^S_{jj})(\mu_{jp}-f_j) - \sum_{l\neq j, l\in S} \kappa^S_{jl}(\mu_{lp}-f_l) \right] \\ &= \tilde{\gamma}_p(1-\bar{\kappa}_{jj})(\mu_{jp}-f_j) - \tilde{\gamma}_p \sum_{l\neq j} \sum_{S\in\mathcal{S}_j} \kappa^S_{jl} \mathbf{1}\{l\in S\} \frac{\phi(S;\boldsymbol{\theta})}{\phi_j(\boldsymbol{\theta})} (\mu_{lp}-f_l) \\ &= \tilde{\gamma}_p(1-\bar{\kappa}_{jj})(\mu_{jp}-f_j) - \tilde{\gamma}_p \sum_{l\neq j} \phi_l \mathbb{E}[\bar{\kappa}^S_{jl}|j,l\in S](\mu_{lp}-f_l) \\ &= \tilde{\gamma}[I-\bar{\mathcal{K}}]'_j(\boldsymbol{\mu}_p - \boldsymbol{f}) \end{aligned}$$

with

$$\bar{\kappa}_{jl} \equiv \begin{cases} \mathbb{E}[\bar{\kappa}^S_{jj}|j\in S] & \text{if } j=l \\ \phi_l(\boldsymbol{f};\boldsymbol{\theta})\mathbb{E}[\bar{\kappa}^S_{jl}|j,l\in S] & \text{if } j\neq l. \end{cases}$$

Overall, fund j pricing problem can be written more compactly as

$$\max_{f_j} \quad P\cdot(f_j-c_j)\cdot\phi_j(\boldsymbol{f}_j;\boldsymbol{\theta})\cdot[I-\bar{\mathcal{K}}]'_j(\boldsymbol{\mu}-\boldsymbol{f})\cdot\tilde{\gamma}\bar{A} \tag{2.69}$$

where for simplicity I assumed that investors preferences are homogeneous too.[53]

In what follows, I first prove a lemma that provides conditions ensuring that funds' problem is concave and then show that funds' objective satisfies such conditions under the previous assumptions. Equilibrium existence will then follow directly from Kakutani fixed point theorem.

Lemma 1. *Let $\pi(f)$ be a continuous and strictly concave function $\pi''(f)<0$ that is uniquely maximized at f^* and is such that $\pi(f^*)>0$. Let $\phi(f)$ a continuous and decreasing function*

53. This assumption is irrelevant for the existence result.

such that

$$\phi(f) \in (0,1) \tag{2.70}$$

$$\phi'(f) = -\phi(f)(1-\phi(f)) < 0 \tag{2.71}$$

then the function $\phi(f)\pi(f)$ *is concave on a compact set* $[\underline{f}, \overline{f}]$ *with unique interior maximizer* $f^{**} \leq f^*$.

Proof: Because π is continuous and $\pi(f^*) > 0$, we can construct $[\underline{f}, \overline{f}]$ such that $\pi(f) > 0$ for all $f \in [\underline{f}, \overline{f}]$ and $f^* \in [\underline{f}, \overline{f}]$.

Next suppose there exists a $f^{**} \in (\underline{f}, \overline{f})$ such that

$$\phi'(f^{**})\pi(f^{**}) + \phi(f^{**})\pi'(f^{**}) = 0 \tag{2.72}$$

which can be rearranged as

$$-\frac{\phi'(f^{**})}{\phi(f^{**})} = \frac{\pi'(f^{**})}{\pi(f^{**})} \tag{2.73}$$

Taking the second order condition

$$\phi''(f^{**})\pi(f^{**}) + 2\phi'(f^*)\pi'(f^*) + \phi(f^{**})\pi''(f^{**}) < 0 \tag{2.74}$$

The last term of the above is negative by assumption. The sum of the first two terms is also

negative:

$$\phi''(f^{**})\pi(f^{**}) + 2\phi'(f^{*})\pi'(f^{*}) < 0 \tag{2.75}$$

$$\Leftrightarrow \frac{\phi''}{\phi'} + 2\frac{\pi'}{\pi} > 0 \tag{2.76}$$

$$\Leftrightarrow \frac{\phi''}{\phi'} - 2\frac{\phi'}{\phi} > 0 \tag{2.77}$$

$$\Leftrightarrow -(1-2\phi) + 2(1-\phi) = 1 > 0 \tag{2.78}$$

where the latter equivalence uses the fact that $\phi'' = -\phi'(1-2\phi)$. This shows that if an interior f^{**} that satisfies the necessary FOC exists then it is always a maximum which implies that $\phi(f)\pi(f)$ is concave on $[\underline{f}, \overline{f}]$.

Lastly, we can show that such p^{**} indeed exists and is interior. To see this evaluate the FOC at $\underline{f}$ and note that under the above assumptions the following

$$\phi'(\underline{f})\pi(\underline{f}) + \phi(\underline{f})\pi'(\underline{f}) > 0 \tag{2.79}$$

holds by choosing $\underline{f}$ sufficiently low (for instance choosing $\underline{f}$ = marginal cost such that $\pi(\underline{f}) = 0$) and by noting that $\pi'(\underline{f}) > 0$ because $\underline{f} < f^{*}$. Then evaluate the FOC at $\overline{f}$ and note that

$$\phi'(\overline{f})\pi(\overline{f}) + \phi(\overline{f})\pi'(\overline{f}) < 0 \tag{2.80}$$

which implies, by continuity that there exists an interior f^{**} that satisfies the FOC. Moreover, it must be the case that $f^{**} < f^{*}$.

Overall, conditions all conditions of the lemma hold and we can be assured that fund j objective in (2.69) is concave in f_j which ensures that funds' best replies $f_j(f_{-j})$ are proper functions. Thus, all conditions of Kakutani fixed point theorem are satisfied and a Nash

equilbrium exists (see MWG chp. 8 Proposition 8.D.3).

Funds' maximization problem in (2.69) satisfies the conditions in the previous lemma after defining $\pi(f_j) \equiv (f_j - c_j) \cdot [I - \bar{\mathcal{K}}]'_j(\boldsymbol{\mu} - \boldsymbol{f})$. First note that, because funds' believe that sponsors include at most one fund per category we have that

$$\phi_j(f_j) = \lambda_g \phi_j^1 = \frac{\exp(V_j(\boldsymbol{\theta}))}{1 + \sum_l \exp(V_l(\boldsymbol{\theta}))}$$

which is decreasing in f_j and such that

$$\phi'_j = -\theta_f \phi_j (1 - \phi_j) < 0.$$

Next, note that because funds' believe that $q = 1$, the matrix $\mathcal{K}$ does not depend on f_j as I showed in previous derivations. This in turn implies that $\pi(f_j)$ is quadratic in f_j and thus concave f_j if and only if

$$(1 - \bar{\kappa}_{jj}) > 0.$$

The latter holds because $\bar{\kappa}_{jj} = \mathbb{E}[\kappa_{jj}^S | j, l \in S]$ and $\kappa_{jj}^S \in (0, 1)$ for any S as I showed in previous derivations. Because π is globally concave in f_j it admits a unique maximum, f_j^*. Moreover we have $f_j^* > c_j$ whenever the previous dominance diagonal holds:

$$(1 - \bar{\kappa}_j)(\mu_j - c_j) > \sum_{l \neq j} |\bar{\kappa}_{jl}|(\mu_l - c_l) \tag{2.81}$$

and $f_l \in [c_l, \mu_l]$ for all l. To see this note that the above condition implies that

$$\begin{aligned}(1-\bar{\kappa}_j)(\mu_j - c_j) &> \sum_{l\neq j} |\bar{\kappa}_{jl}|(\mu_l - c_l) \\ &\geq \sum_{l\neq j} |\bar{\kappa}_{jl}|(\mu_l - f_l) \\ &> \sum_{l\neq j} \bar{\kappa}_{jl}(\mu_l - f_l)\end{aligned}$$

where the last two inequalities holds whenever $f_l \in [c_l, \mu_l]$. But note that the previous inequality corresponds to fund j FOC (when maximizing π) evaluated at $f_j = c_j$

$$\begin{aligned}\frac{\partial \pi(f_j)}{\partial f_j} &= (1-\bar{\kappa}_{jj})(\mu_j - f_j) - \sum_{l\neq j} \bar{\kappa}_{jl}(\mu_l - f_l) - (1-\bar{\kappa}_{jj})(f_j - c_j)\Big|_{f_j=c_j} \\ &= (1-\bar{\kappa}_j)(\mu_j - c_j) - \sum_{l\neq j} \bar{\kappa}_{jl}(\mu_l - f_l) > 0.\end{aligned}$$

Then it must also be the case that $\pi(f_j^*) > 0$, if not setting $f_j = c_j$ would lead to higher π which would be a contradiction.

Microfoundation of the distribution of the number of options. In what follows I offer a simple microfoundation for the distribution of the number of options sponsors include in any given category building on the Stigler (1961) simultaneous search model.

Sponsor p first commits to include n investment options in investment category g and then conditional on n, selects the n options providing her with the n highest utilities. I assume that sponsors choose n before observing their random utility shocks ε_{jp} but knowing the mean utility of each option $V_j(\boldsymbol{\theta}_p)$. From this perspective the utility sponsor p derives from option j is distributed as

$$u_{jp} \sim T1EV(V_j). \tag{2.82}$$

I assume that sponsors incur a cost $c(n)$ for including n options which is increasing, convex in n and is such that $c(1) < \mathbb{E}[u_{j_1 p}]$. The benefit from choosing n options is given by

$$\sum_{s=1}^{n} \mathbb{E}[u_{j_s p}] \tag{2.83}$$

where j_s is the option that provides the sth highest utility. Overall, sponsor p problem becomes

$$\max_{n \geq 1} \sum_{s=1}^{n} \mathbb{E}[u_{j_s p}] - c(n). \tag{2.84}$$

The are two main differences between this model and the Stigler (1961) model. First, in this case the agent commits to consume n options whereas in Stigler (1961) the agent commits to sample n options and among those to consume the one with the highest utility. Second, in this model after committing to n, sponsor p observes the realized utility of all options available but has chosen to only consume the n highest whereas in the Stigler (1961) model the agent observes the utility realizations of the searched options only, and among those selects the highest.

The solution to the above problem is given by the n^* such that the marginal benefit from chosing to include $n^* + 1$ options is lower than the change in the cost

$$\mathbb{E}[u_{j_{n^*+1}}] \leq c(n^* + 1) - c(n^*). \tag{2.85}$$

Heterogeneity in the cost of adding options c_p or in the benefits u_{jp} across sponsors would produce a different n_p^* for each sponsors. In the data such distribution of n^* corresponds to the one plotted in Figure 2.9 suggesting that most sponsors do not include more than one option and that the likelihood decreases geometrically in the number of options.

In estimation I do not attempt to estimate the distribution of costs c_p that matches the

observed distribution in Figure 2.9 because it would complicate substantially the estimation of sponsor preferences. First, it would require estimating such distribution at each iteration of the estimation algorithm because the optimal number of options n^* itself depends on sponsors preference parameters. Second, it would require finding a solution to problem (2.84) which is non-convex without further restrictions.

To keep estimation tractable I instead model sponsors' choice of n as a random draw from the empirical distribution which I parametrize as geometric with parameter q. In estimation I allow for such distribution to be heterogeneous at the recordkeeper-year-category level. In general, the distribution of the number of options included within each category looks similar to the one in Figure 2.9 for many cuts of the data I have considered.

Derivation of investor surplus. Consider active investor i in plan p. Investor i has preferences over its retirement portfolio allocation $\boldsymbol{a}_i$ given by

$$u_i(\boldsymbol{a}_i) = \boldsymbol{a}_i'(\boldsymbol{\mu} - \boldsymbol{f}) - \frac{\gamma}{2}\boldsymbol{a}_i' V \boldsymbol{a}_i$$

where

$$V \equiv I + X_{(2)}X_{(2)}'.$$

Investor i's demand implied by the previous problem is

$$\boldsymbol{a}_i(\boldsymbol{f}) = \frac{1}{\gamma}V^{-1}(\boldsymbol{\mu} - \boldsymbol{f}).$$

Next, combine the expression for $\boldsymbol{a}_i$ with the expression for $u_i(\boldsymbol{a}_i)$ as follows

$$\begin{aligned} u_i(\boldsymbol{a}_i) &= \boldsymbol{a}_i'(\boldsymbol{\mu} - \boldsymbol{f}) - \frac{\gamma}{2}\boldsymbol{a}_i' V \boldsymbol{a}_i \\ &= \boldsymbol{a}_i(\boldsymbol{f})'(\boldsymbol{\mu} - \boldsymbol{f}) - \frac{1}{2}\boldsymbol{a}_i(\boldsymbol{f})'(\boldsymbol{\mu} - \boldsymbol{f}) \\ &= \frac{1}{2}\boldsymbol{a}_i(\boldsymbol{f})'(\boldsymbol{\mu} - \boldsymbol{f}), \end{aligned}$$

which is the measure of investors' surplus provided in the main text.

Example of category-based correlation structure. Consider a plan menu with 8 assets classified in the following four investment categories 'Equity-Growth', 'Equity-Value', 'Bond-Government' and 'Bond-Corporate'. Also assume that there are two assets for each of the four categories.

The vector of characteristic for a given asset j, $\tilde{\boldsymbol{g}}_j$, has six elements corresponding to the 1st level characteristics (Equity, Bond) and 2nd level characteristics (Equity-Growth, Equity-Value, Bond-Gov, Bond-Corp). If asset j is an Equity-Value fund its vector of characteristics is given by:

$$\tilde{\boldsymbol{g}}_j = (1, 0, 0, 1, 0, 0)'. \tag{2.86}$$

With the assumption that there are two assets within each category the 8×8 outer-

product matrix G_p is given by:

$$G_p = \tilde{G}_p \tilde{G}_p' = \begin{bmatrix} 2 & 2 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 2 & 2 \end{bmatrix}$$

Funds' investment category classifications capture cross-substitution patterns between assets. In this example investors treat Equity and Bond assets as independent and within Equity, Growth and Value assets as less substitutable than the two Growth or the two Value assets.

2.12 Appendix: Turnover ratios & identification

To consistently estimate sponsors' and investors' preferences I instrument funds' fees with funds' turnover ratios which capture trading costs that are typically pass on to investors through fees. In this appendix, I first provide more details on how a fund's turnover is computed and, after that, I discuss a possible threat to identification. Lastly, to motivate the relevance of the instrument, I provide an illustrative example of how funds' turnover might affect funds' fees.

2.12.1 Funds' turnover ratios.

I obtain data on funds' turnover from CRSP, which reports it at fiscal year frequency. Turnover for fund j in year t is defined as

$$\text{turnover}_{jt} = \frac{\min(\text{buys}_{jt}, \text{sells}_{jt})}{\text{Average TNA}_{jt}}$$

where the numerator is the smaller of the funds' total purchases and sales over fiscal year t and the denominator is the average total net asset value (TNA) in year t. This measure is the one that the SEC requires funds' to report each year.

A key advantage of this measure is that by taking the minimum between sales and purchases it excludes turnover arising from trading activities triggered by persistent inflows or outflows. For example, if a fund experiences substantial inflows most of the trading activity will be implemented to buy more securities or increase current portfolio positions. In this case though, because of the min() in the numerator, the turnover reported will capture the sales and not the purchases. Similarly, if a fund experiences substantial outflows, the turnover measure will likely pick up the fund's purchases. Overall, because flows are notoriously persistent, this measure of turnover is largely immune to flows and instead will capture discretionary trading decision from funds' managers (Pástor, Stambaugh and Taylor (2017)).

From an identification perspective, this property of the turnover measure is particularly appealing because it makes it mechanically less dependent of persistent demand shocks. Nonetheless, the measure is not completely independent of demand shocks that generate non-persistent inflows or outflow and, as such, may be a non valid instrument if it correlates with some driver of funds' flows.

2.12.2 Identification: turnover vs. performance

Funds' performance is a well-known driver of funds' flows (Chevalier and Ellison (1997)) and recent research suggests that active funds achieve better performance when they trade more and have higher turnover ratio (Pástor, Stambaugh and Taylor (2017)).[54] If investors' chase performance and turnover determines or is correlated with performance then the exclusion restriction I employ to identify sponsors' preferences would be violated.

To reduce the concern about this potential identification threat I check whether my instrument i.e., the residual turnover after absorbing funds' brand, year, category and passive fixed effects, shows significant correlation with some measures of investment performance and find that this is not the case.

Table 2.17 presents a correlation table between the residualized turnover (i.e., the instrument), the residualized expense ratio (i.e., the endogenous variable) and three measures of investment performance; (i) a fund's (gross of fees) alpha from a three Fama-French factor regression plus Momentum, (ii) a BrightScope-category-adjusted measure of performance computed as the difference between a fund's gross return and the return of the corresponding BrightScope category and (iii) a Morningstar-category-adjusted measure of performance computed as the difference between a fund's gross return and the return of the corresponding

54. The turnover-performance relationship is non-significant for passive funds. Also, for active funds, the relationship is stronger over time rather than across funds.

	turnover ratio	expense ratio	alpha	MS adj. return	BS adj. return
turnover ratio	1.00	0.16	-0.00	0.00	-0.02
expense ratio	0.16	1.00	0.01	-0.01	0.01
alpha	-0.00	0.01	1.00	0.49	0.54
MS adj. return	0.00	-0.01	0.49	1.00	0.69
BS adj. return	-0.02	0.01	0.54	0.69	1.00

Table 2.17: Correlation table of instrument (turnover ratio) with fund performance measures. Turnover ratio and expense ratios are residuals after absorbing funds' brand, year and category fixed effects. Performance measures are yearly-demeaned.

Morningstar-category-adjusted.

Turnover and expense ratio exhibit a positive correlation confirming that the instrument has a strong first stage. Conversely, turnover does not seem to be correlated with any of the performance measure considered, alleviating the concern of a possible violation of the exclusion restriction.

To check the statistical significance of these correlations I present a series of binscatter plots where I regress the instrument on the above mentioned measures of performance and on the endogenous variable. Figure 2.25 shows the strong and significant correlation between turnover and expense ratio. The other set of figures (2.26, 2.27, 2.28) instead show that there is no significant correlation between the instrument and investment performance. Lastly, in Figures (2.29, 2.30, 2.31) I check whether turnover correlates with performance in the following year and still find no evidence of a significant correlation, again alleviating the concern of a possible violation of the exclusion restriction.

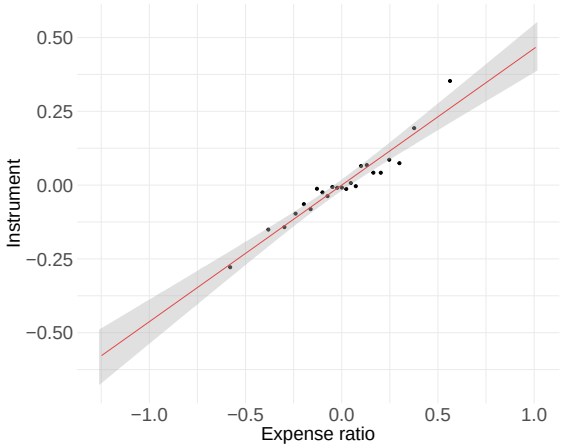

Figure 2.25: Expense ratio is residual expense ratio after absorbing funds' brand, year and category fixed effects. Instrument is residual turnover ratio.

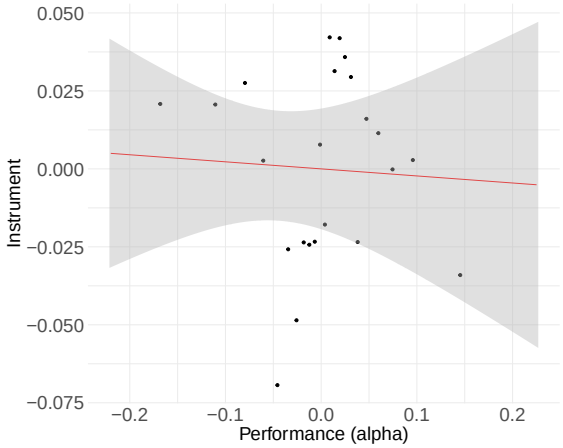

Figure 2.26: Performance is yearly-demeaned alpha from 3 FF factors plus Momentum. Instrument is residual turnover.

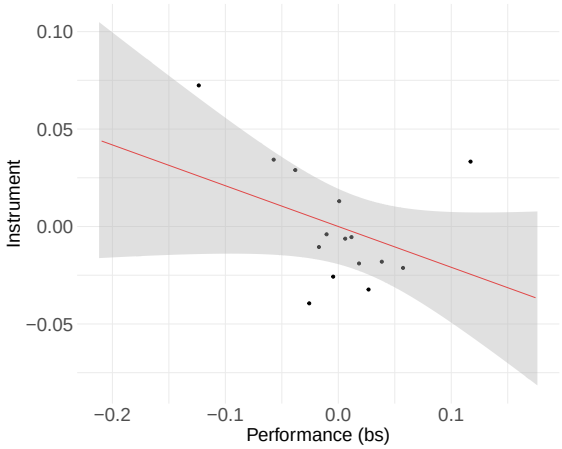

Figure 2.27: Performance is yearly-demeaned (gross) return relative to BrightScope category return. Instrument is residual turnover.

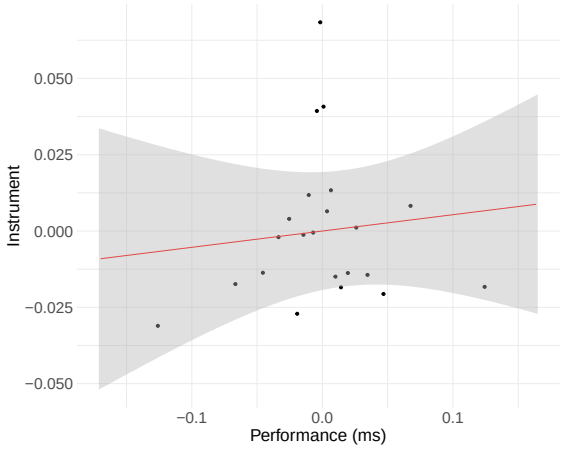

Figure 2.28: Performance is yearly-demeaned (gross) return relative to Morning Star category return. Instrument is residual turnover.

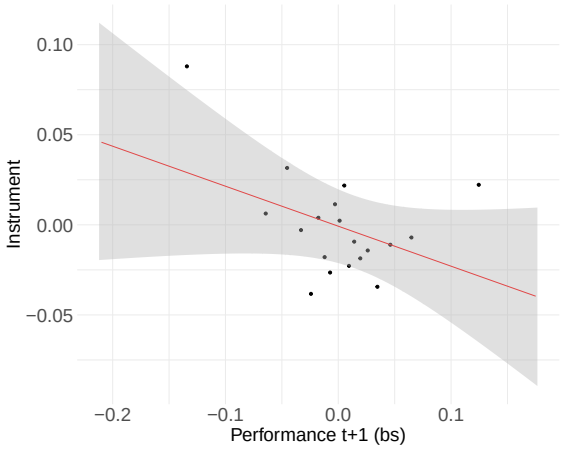

Figure 2.29: Performance is next period yearly-demeaned (gross) return relative to BrightScope category return. Instrument is residual turnover.

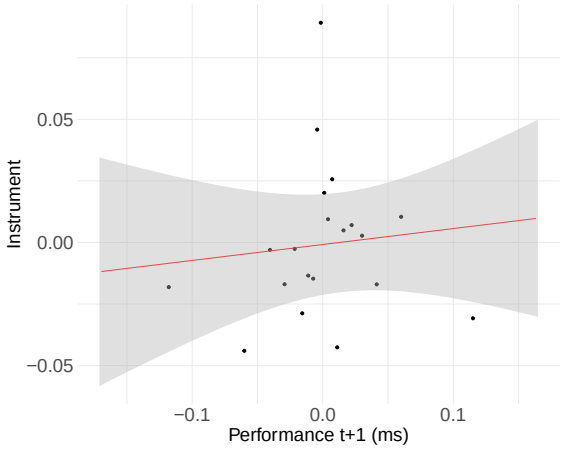

Figure 2.30: Performance is next period yearly-demeaned (gross) return relative to Morning Star category return. Instrument is residual turnover.

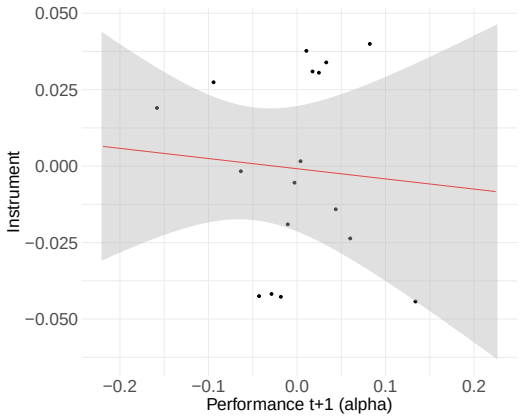

Figure 2.31: Performance is next period yearly-demeaned alpha from 3 FF factors plus Momentum. Instrument is residual turnover.

2.12.3 How does turnover affect fees?

Figure (2.25) provides evidence of a strong relationship between turnover (i.e., the instrument) and fees (i.e., the endogenous variable). In this section, I show how a simple model in which funds' maximize their fee revenue requires fund managers to pass trading costs onto investors via higher fees.

Following Pástor, Stambaugh and Taylor (2020), consider fund j who optimally chooses its fees f_j to maximize its dollar profit:

$$\max_{f_j}(f_j - c_j(t_j))A_j(f_j, t_j)$$

where A_j is the dollar AUM of the fund, c_j is the marginal cost of operating the fund and t_j is the fund turnover. In a world in which investors chase performance it is reasonable to assume that a fund's AUM depend on the level of fees and turnover. For example, when investors are rational and supply capital perfectly elastically a la Berk and Green (2004), the AUM of fund j must be such that investors expect zero returns net of fees and trading costs:

$$\mu_j(t_j) - f_j - q_j(A_j, t_j) = 0 \tag{2.87}$$

where μ_j is fund j expected return gross of fees and q_j is a per-dollar trading cost. The return generated by the fund may depend on its turnover if, for instance, higher skilled managers trade more (Pástor, Stambaugh and Taylor (2017)). A fund trading cost might depend on its size (Berk and Green (2004)) and on its portfolio turnover (Pástor, Stambaugh and Taylor (2020)). Overall, investors' behaviour (i.e., equation (2.87)) will determine fund j's demand $A_j(f_j, t_j)$ as a function of fees and turnover.

Funds' turnover ratio can also affect its operating costs. For example, funds that trade more might incur in higher operating costs (e.g., higher more research analysts) and may

need to implement more rewarding compensation structures for their portfolio managers which force the fund to charge higher advisory fees to its investors (Ma, Tang and Gomez (2019)).

The optimal fee charged by fund j will then depend on a fund turnover or expected turnover t_j

$$f_j = c_j(t_j) + \frac{A_j(f_j, t_j)}{\partial A_j(f_j, t_j)/\partial f_j}.$$

Overall, funds' optimizing behaviour explains why we observe a positive relationship between trading costs and fees (Pástor, Stambaugh and Taylor (2020)). Funds' managers pass trading costs to investors to maximize dollar profits which provides a rationale for using turnover as a cost-shifter instrument for fees.

2.13 Appendix: Nesting investors & sponsors preferences

In Section 2.4 I assumed that sponsors have their own preference parameters over funds' attributes. When making their plan inclusion decisions sponsors evaluate the suitability of each fund based of their preference parameters which I denoted by $\boldsymbol{\theta}_p$. In this appendix I illustrate how sponsors' estimated parameters can be interpreted as a weighted average of what I refer to as sponsors' 'true' preference parameters, denoted by $\boldsymbol{\theta}_p^s$ and investors preference parameters denoted by $\boldsymbol{\theta}_p^i$.

To this end, suppose sponsor p random utility from including fund j is given by

$$u_{jp} = \chi V_{jp}^{\text{sponsor}}(\boldsymbol{\theta}_p^s) + (1-\chi) V_{jp}^{\text{investor}}(\boldsymbol{\theta}_p^i) + \zeta_j + \varepsilon_{jp}$$

where the weight $(1-\chi) \in (0,1)$ captures how much sponsors account for investors preferences when making their plan inclusion decisions.

As before, let us assume that sponsors mean utility $V_{jp}^{\text{sponsor}}(\boldsymbol{\theta}_p^s)$ is linear in funds' characteristics, formally $V_{jp}^{\text{sponsor}} = \boldsymbol{w}_{jp}' \boldsymbol{\theta}_p^s$. The parameter vector $\boldsymbol{\theta}_p^s$ captures how much a non-altruistic sponsor (e.g., $\chi = 1$) would value funds' attributes. For example, a non-benevolent sponsor might prefer to include expensive funds (i.e., its 'true' fee parameter is positive $\theta_{pf}^s > 0$) to maximize its recordkeeper revenues from indirect compensations (Bhattacharya and Illannes (2022)). If sponsors internalize their investors utility then sponsors' parameters estimated starting from the preference specification in (2.1) will also reflect part of investors preferences. To see this suppose that investors mean utility $V_{jp}^{\text{investor}}(\boldsymbol{\theta}_p^i)$ is also linear in funds' characteristics i.e., $V_{jp}^{\text{investor}}(\boldsymbol{\theta}_p^i) = \boldsymbol{w}_{jp}' \boldsymbol{\theta}_p^i$. Then, it is straightforward to see that the parameter $\boldsymbol{\theta}_p$ we estimated in the main text is a weighted average of sponsors' and

investors' preferences

$$u_{jp} = \boldsymbol{w}'_{jp} \underbrace{(\chi \boldsymbol{\theta}_p^s + (1-\chi)\boldsymbol{\theta}_p^i)}_{\equiv \boldsymbol{\theta}_p} + \zeta_j + \varepsilon_{jp}. \tag{2.88}$$

At this point there are two challenges. First, investors' indirect utility is more complex than the simple linear specification I used just above. Second, it is unclear how we can identify and estimate χ. In what follows, I explain why I abstract from the first challenge and, under the assumption that investors utility is linear, I will illustrate what restrictions are needed to identify χ and I will provide an estimate of it.

Assuming that investors indirect utility is linear in funds' characteristics is not consistent with the quadratic type of preferences I specified when defining investors' portfolio problem. Under those type of preferences investors indirect utility depends on the characteristics of all the funds included in the plan menu because investors optimally diversify across the funds available. In other words, investors indirect utility depends on the menu sponsors choose on their behalf. For this reason, nesting investors' utility into sponsors random utility is not straightforward because it would require sponsors taking expectation over all possible menus that could be chosen. Thus, I will take a different approach and specify the utility investors derive from having fund j in their menu as a linear function of j characteristics. This would be consistent with a demand model, like Berry (1994), in which plan investors make a discrete choice among the options available in their plan, which under the appropriate coefficient restrictions, would be observationally equivalent to the asset demand derived from mean-variance preferences (Koijen and Yogo (2019)).

With the assumption that V_{jp}^{investor} is linear in funds characteristics we can identify and estimate χ as follows. First, we need to obtain an estimate of investors' preferences $\hat{\boldsymbol{\theta}}_p^i$ and an estimate of $\hat{\boldsymbol{\theta}}_p$. Second we need to assume that at least one fund characteristic enters investors' preferences but is excluded from sponsors' preferences. For example, if investors

	Preference parameters	
	Active Investors	All Investors
χ	0.749	0.749
Expense ratio (f, bp.)		
θ_{pf}	-0.021	-0.021
θ^i_{pf}	-0.039	-0.025
θ^s_{pf}	-0.019	-0.019
Affiliation (a)		
θ_{pa}	0.823	0.823
θ^i_{pa}	0.276	0.174
θ^s_{pa}	1.040	1.040
Return (r, pp.)		
θ_{pr}	0.004	0.004
θ^i_{pr}	0.025	0.016
θ^s_{pr}	0	0

Table 2.18: Estimated investors and sponsors parameters assuming investors make their portfolio choice according to a discrete choice demand with linear random utility.

care about past returns but sponsors do not (i.e., $\theta^s_{pr} = 0$) then χ will be determined by

$$\chi = 1 - \frac{\hat{\theta}_{pr}}{\hat{\theta}^i_{pr}}.$$

Table 2.18 presents the estimates of sponsors and investors preferences assuming that active investors solve a discrete choice portfolio problem and that past returns gross of fees enter investors preferences but not sponsor preferences. Under this assumption investors' preference parameters $\boldsymbol{\theta}^i$ can be estimated from the following linear regression:

$$\log(s^{\text{active}}_{jpt}) = \boldsymbol{w}'_{jpt}\boldsymbol{\theta}^{i,\text{active}} + \psi_{pt} + \xi_{jpt} \tag{2.89}$$

where s_{jpt}^{active} is the active investors portfolio share of fund j in plan p and year t, ξ_{jpt} is an unobserved demand shock possibly correlated with fees and ψ_{pt} are plan-by-year fixed effects which are needed to absorb the inclusive value component of investors discrete choice problem.[55] The data does not distinguish between portfolio shares of inactive v. active investors. To recover the latter I use the estimated fraction of inactive investors and obtain the portfolio share of the active as

$$s_{jp}^{\text{active}} = s_{jp}\mathbf{1}\{j \neq d\} + \frac{s_{jd} - \delta_d}{1 - \delta_d}\mathbf{1}\{j = d\}.$$

To account for the endogeneity of fees I use funds' turnover ratio as instrument. The parameters $\boldsymbol{\theta}_p$ are estimated from the menu choice problem I developed in Section 2.6. Specifically, I use the estimates reported in the left column of Table (2.9). Assuming inactive investor preference parameters is equal to 0 I can recover the preference weighting as

$$\chi = 1 - \frac{\hat{\theta}_{pr}}{\hat{\theta}^i_{pr}(1 - \delta)} \tag{2.90}$$

The estimates suggest that preference misalignment are important in this market. Sponsors weight their own preferences roughly three times more (0.75/0.25) than their investors preferences. Looking at the estimated coefficients, it appears that the largest misalignment is in the preference for fund affiliation. Furthermore, consistent with previous results, sponsors tend to tolerate higher fee more than their investors and particularly so if compared to active investors. Active investors marginal dis-utility from higher fees is twice larger than the one of their sponsors.

55. Plan-by-year fixed effect absorb the following term $\ln(1 + \sum_{j \in S_{pt}} \exp(V_{jpt}^{\text{investors}}))$ which is plan-year specific. Logit demand implies is given by $s_{jpt} = \frac{\exp(V_{jpt}^{\text{investors}})}{1 + \sum_{j' \in S_{pt}} \exp(V_{j'pt}^{\text{investors}})}$.

2.14 Appendix: 401(k) Lawsuits

This appendix provides details about some recent 401(k) lawsuits. I start by providing some background on 401(k) regulations building on (Mellman and Sanzenbacher (2018)) and then offer few specific examples of recent lawsuits.

The design of 401(k) retirement plans is governed by the Employee Retirement Income Security Act of 1974 (ERISA), with the Department of Labor (DOL) in charge of updating and enforcing such regulation. The law specifies that plan sponsors (i.e., employers) have a fiduciary duty to their plan investors requiring them to design and administer the plan in the 'sole benefit' of plan participants.

While the regulation is clear about the role of plan sponsors as fiduciaries, it provides almost no guidance on how to fulfill such duty in practice. For example, not much is said about how plan fiduciaries should select the type and number of investment options or determine a reasonable level of fees. Instead of laying out specific regulations or guidance, the DOL's general approach to overseeing 401(k)s has been through its own enforcement actions or through privately initiated litigation. Overall, plan fiduciaries are often left to guess what practices comply with ERISA and may only become aware of an alleged violation from a DOL investigation or lawsuit.

Typically there are two reasons that trigger 401(k) lawsuits. First, the inclusion of inappropriate investment options and second, the inclusion of options charging excessive fees. The former was the most common cause after the Great Recession mainly as a consequence of the inclusion of poor-performing employers' own stock. However, this kind of lawsuit has become less common since a 2014 Supreme Court ruling in the case of *Dudenhoeffer v. Fifth Third Bancorp* indicating that plan fiduciaries will not be held liable for failure to predict the future performance of the employers stock. Since then, most lawsuits involved allegations of excessive investment and administrative fees. In what follows, I describe a few recent examples.

Allen v. M&T Bank Corp (2016). The Plaintiff (Allen) alleges that the Defendant (M&T) breached their fiduciary duties by retaining their proprietary funds within the plan despite the availability of similar lower cost and better performing investment options. According to the plans Form 5500 filed for 2010, of the 22 mutual fund investment options in the Plan, 8 were from proprietary M&T mutual funds, representing over 30% of all mutual fund investments. However, these proprietary mutual funds charged significantly higher fees than average for performance that most often trailed both the Fund benchmarks and the mutual fund averages.

The Plaintiff provides some specific examples. For instance, the Wilmington Large Cap Value Institutional lagged the performance of a more reasonably priced alternative, Vanguard Equity Income Fund Admiral Shares. The Wilmington fund charged an expense ratio of 1.17%, higher than the Large Cap Value average of 0.83% and the 0.21% fee charged by the Vanguard Equity Income Fund Admiral Shares. A similar observation is made for the Wilmington Funds Small Cap Growth Institutional Fund charging an expense ratio of 1.39% against the 0.40% fee charged by Vanguard Stragic Small Cap Equity Fund.

Creamer v. Starwood Hotels & Resorts Worldwide Inc (2016). The Plaintiff (Creamer) alleges that the Defendant (Starwood Hotels & Resorts Worldwide, Inc.) serially breached its fiduciary duties in the management, operation and administration of its employees 401(k) plan. It failed to ensure that fees charged to participants were reasonable. It caused plan participants who invested in index funds to pay seven times more than a reasonable fee. Indeed, the Starwood Plan received from the BrightScope rating service a score of only 61. The top BrightScope rating for peer plans was 90. The Plaintiff highlights that this difference would require sixteen years of additional by Starwood employees to reach the same level of savings as peer plan participants. Starwood participants lost savings of $110,871 per participant.

The Plaintiff also provides some specific examples. For instance, the BlackRock LifePath

Index funds (the plan TDF) just hold other BlackRock index funds. BlackRock Life Path 2050 Index Fund institutional shares have net operating expenses of 0.20%. The 2050 Index Fund is a fund that invests all of its assets in other BlackRock funds. 52% of the Life Path Index Fund was invested in the BlackRock Russell 1000 Index Fund. The Russell 1000 Index fund had net operating expenses of 0.08%. Thus, the fee paid by plan participants is 0.20% plus 0.08% for a total of 0.28%. In contrast, the Vanguard Institutional Index Fund Institutional Shares had a total expense ratio of only 0.04% so the plan has chosen funds with fees that are 700% more than the comparable Vanguard fund - a difference of 24 basis points

McCorvey v. Nordstrom, Inc. (2017). The Plaintiff (McCorvey) alleges that the Defendant (Nordstrom) failed to adequately and prudently manage the plan. It allowed unreasonable fees to be incurred by participants and failed to use lower cost investment vehicles. The annual operating fees charged for many of the plans investment options were substantially higher than reasonable management and operating fees of comparable funds, both index and actively managed funds. These fees were up to 16 times higher than comparable index funds, and up to 2.7 times higher than comparable actively managed funds.

The Plaintiff highlights that the high fee funds in the Nordstrom plan could have been easily replaced by lower cost index funds, TDF, or actively managed funds. For example, the PIMCO Total Return charging 46 basis points in fees could have been replaced by the Vanguard High Dividend Yield Index Fund charging 15 basis points or the Vanguard Growth and Income Fund Admiral Shares (an active fund) charging 23 basis points. Similarly, the average expense ratio for the set of TDFs available in Nordstrom (42 basis points) could have been five times lower by replacing it with Vanguard Institutional Target Date Funds whose average expense was around 9 basis points.

Pledger v. Reliance Trust (2015). The Plaintiff (Pledger) alleges that the Defendant

(Reliance Trust) breached its fiduciary by providing to the plan investment options that contained unreasonable management fees when cheaper versions of the same investments were available to the plan, as were other high-quality, low-cost institutional alternatives. The Plaintiff also alleges that the plan recordkeeper (Insperity) and the Defendant engaged in self-dealing by offering higher-cost investments to the plan's participants, because Reliance selected those investments in order to pay a larger amount of revenue-sharing to the recordkeeper.

Schapker v. Waddell & Reed Fin. Inc (2017). The Plaintiff (Schapker) alleges that the Defendants (WR Financial) selected the investment opportunities made available to the plan participants. During the Class Period, more than 97% of the investment opportunities made available to the plan participants were established and managed by WR Financial or its affiliates. Only one unaffiliated investment optionout of dozens of funds offered each yearwas ever offered to plan participants. Because nearly all the investment options the Defendants made available to plan participants were established and managed by WR Financial or its affiliates, the Defendants caused the plan to pay its own Sponsor, WR Financial.

Further the Plaintiff points out that the fees charged to plan participants for their investments were in excess of the fees typically charged by unaffiliated companies for comparable mutual funds and products, and the performance levels of the investment options within the plan were worse than the performance achieved by unaffiliated companies for comparable mutual funds and investment products. Defendants could have selected comparable investment products from unaffiliated companies that cost less and performed better than the proprietary branded investment products to which the Defendant limited the plan participants.

CHAPTER 3

OLIGOPOLISTIC COMPETITION, FUND PROLIFERATION, AND ASSET PRICES

3.1 Introduction

The US mutual fund industry has witnessed a tremendous shift from active to passive investing in the past two decades. The share of assets under management (AUM) held in passive equity funds increased from about 20% at the beginning of 2000 to 50% in the first quarter of 2020 (Figure 3.1), making the passive industry the dominant one in the US equity market. While this shift from active to passive is a well-known fact, less is known about the competitive dynamics within the passive industry and their implications for asset prices and investors' welfare.

We start this paper by highlighting two facts about the structure of the passive mutual fund industry. First, the industry is concentrated in a handful of large fund families. Figure 3.2 plots the market share of the five biggest families for three different cap-based categories within the US passive equity industry: Large Cap, Mid Cap and Small Cap. In each category, the market share of the top five families has been roughly steady over time and amounts to more than 80% of the market.

Second, perhaps not surprisingly, the number of passive investment vehicles such as index funds and ETFs have been increasing in the past 20 years. More interestingly, Figure 3.3 shows how the average number of passive funds per fund family evolved over time, separately for the five biggest families (blue line) and the rest of them (red line). The pattern is striking: not only do the five biggest families manage more assets, but they do so by deploying many more funds relative to their competitors. In each investment category, this gap in the number of funds has been increasing over time, suggesting that fund proliferation might be a key

mechanism through which these investment firms compete. To maintain their market shares, the largest families keep introducing new funds to capture household investors' demand.

To rationalize these patterns, we develop and estimate a dynamic oligopoly model of the mutual fund industry. A key component of our model is the distinction between individual mutual funds and fund families which we also refer to as management companies. While there exists an extensive literature that studies the mutual fund industry, most of it focuses on funds and disregards the role of management companies.[1] In practice, management companies are responsible for relevant decisions that shape the competitive environment in which funds operate. They decide if and when to introduce new funds, choose the investment sector in which to operate their new funds and establish the investment style of the new funds they create (e.g., the investment mandates). To accommodate the presence of both management companies and funds, to emphasize the key role of fund initiation and to analyze the competitive forces that drive the dynamics of this industry, our model builds on two layers of Cournot competition. In the inner layer, mutual funds compete by choosing quantities and make a profit from fees (as a percentage of their AUM) determined, in equilibrium, by the investment service demand of a representative household investor. In the outer Cournot layer, an oligopoly of heterogeneous management companies compete with each other by deciding how many funds to operate.

On top of allowing the number of funds to emerge endogenously in equilibrium, the model links the technological primitives of the asset management industry, such as marginal costs and scale economies, to the elasticity of asset demand. We close the model and derive the equilibrium price that clears the asset market under the assumptions of strict mandates and a fixed supply of shares. After setting up the model, in Proposition 7, we analytically prove the existence and uniqueness of a steady state equilibrium in which the number of

1. Some exceptions are Massa (2003), Gaspar, Massa and Matos (2006), Bhattacharya, Lee and Pool (2013), Berk, Binsbergen and Liu (2017), Betermier, Schumacher and Shahrad (2022) and Ørpetveit (2021).

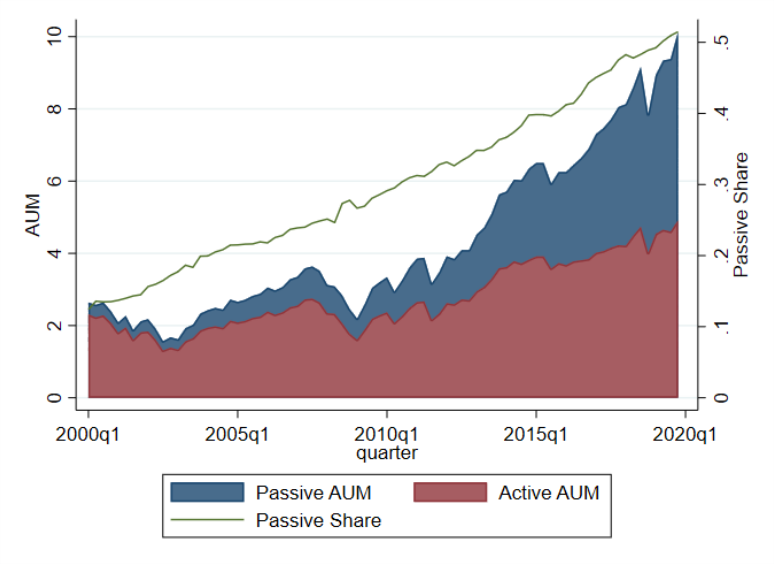

Figure 3.1: Left Axis: AUM in trillions of $ for both passive and active equity industry. Right Axis: Share of AUM held in the passive industry.

funds and the index price are constant. While we characterize the steady-state equilibrium analytically, we further solve the full model numerically. To do so, we develop a nested fixed point numerical routine that, for given parameter values, solves for the optimal equilibrium path to a given terminal condition. At each point in time, the path for the number of funds created is a pure strategy Markov Perfect Nash equilibrium of the dynamic game between management companies, and the path for asset prices clears the asset market in every period.

Secondly, we push forward the recent asset pricing literature that studies the role of institutional investors in determining asset market movements.[2] Contrary to the traditional assumptions that investors are atomistic and that their demand shocks are uncorrelated, this literature documents how asset demand is far from perfectly elastic and how demand shocks affect equilibrium asset prices. Intuitively, the large size of these investors and the presence of specific investment mandates contribute to generating correlated demand shocks, which

2. See for instance, Petajisto (2009) Basak and Pavlova (2013), Koijen and Yogo (2019), Haddad, Huebner and Loualiche (2022) and Pavlova and Sikorskaya (2022).

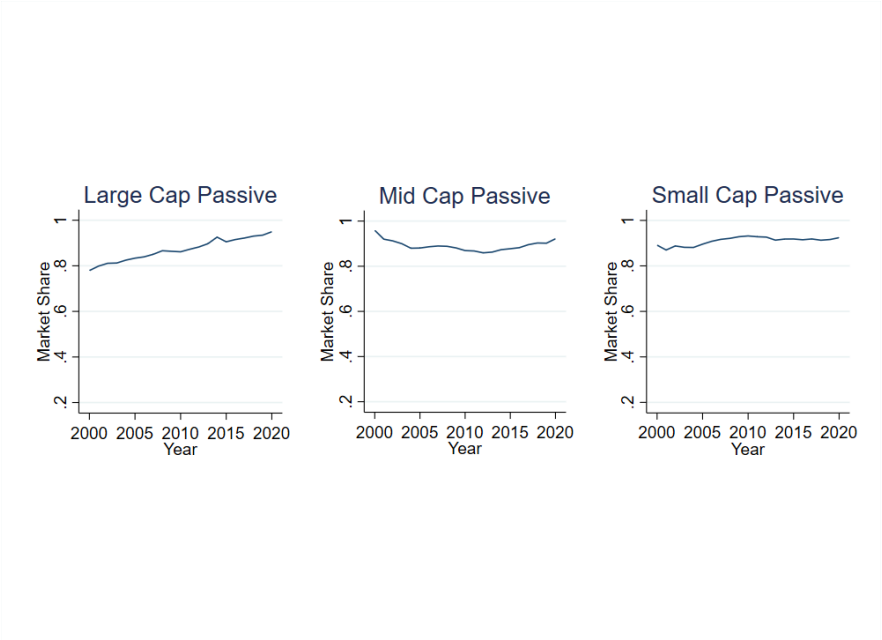

Figure 3.2: Market share of the five biggest investment companies by investment strategy. Market shares are in terms of end-of-year assets under management (AUM).

will inevitably impact asset prices. Our model links the mutual fund industry technology fundamentals to equilibrium asset prices: the price impact of large institutional investors is micro-founded from technology primitives such as fund initiation costs and scale economies. A reduction in initiation costs pushes companies to introduce more products, which will lower the equilibrium fees and, in turn, attract more demand from households. Then, under fixed supply, asset prices will increase to clear the excess demand triggered by the initial reduction in initiation costs.

The second part of the paper turns to the estimation of the model using data on US passive equity funds. We do so by matching two types of data moments, the average per-period fund initiation rate and the average rate at which this fund initiation rate grows. In our model, these moments inform the two parameters that characterize the cost of introducing new funds for management company j, the linear cost parameter c_j and the adjustment cost parameter δ_j. In the data, we compute these moments separately for each of the five

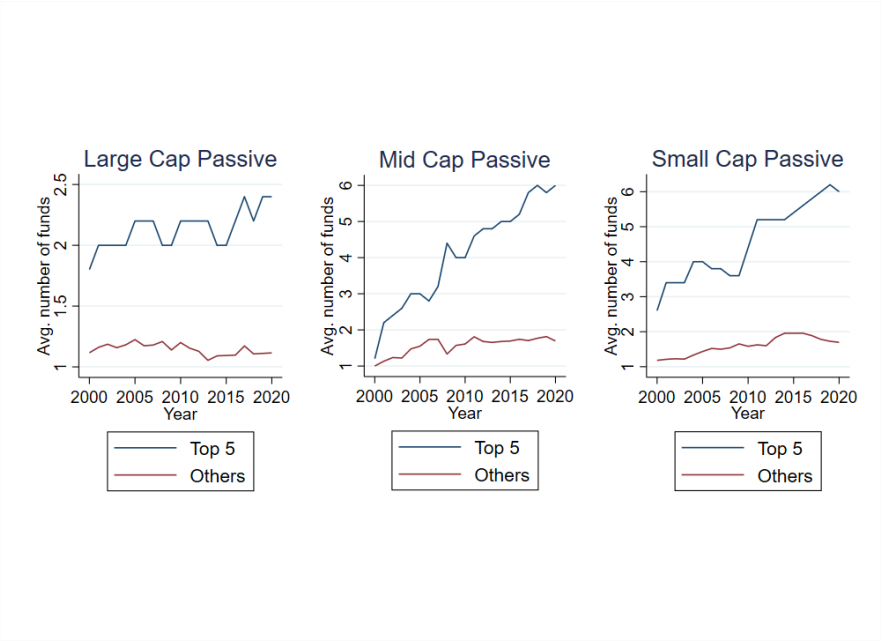

Figure 3.3: Average number of funds per management company. Funds with different share classes count as a single fund.

biggest management companies and the remaining companies pooled together.[3] For each of the five biggest companies, we then obtain an estimate of their fund initiation costs, and we show how our model can match the pattern of fund proliferation observed (and targeted) in our data. As a validation check, we also show how our dynamic model's time series of equilibrium fees closely follows the observed (but untargeted) time series of expense ratios.

Next, with our estimated model, we perform counterfactual exercises to study the welfare effects of removing the largest management companies from the market. We do so for each of the top 5 largest companies Blackrock, Charles Schwab, Fidelity, State Street and Vanguard, one at the time. The welfare effects of removing any one of those companies are large and heterogeneous. We estimate that removing Blackrock or Vanguard from the market would reduce household welfare by 25% and 9.2%, respectively. Such a large welfare

3. Our model only focuses on passive funds, and for the estimation of the model, we will pool all the non-top five companies into one entity which we refer to as the outside company, so that the overall number of companies used in estimation will be 6. As shown in figure (3.12), more than 80% of the market is controlled by the top 5 companies, so the outside group consists of a set of small companies.

loss is a consequence of the fact that although Blackrock and Vanguard are the largest asset managers, they are also the most efficient ones. To further corroborate our finding, in a second set of counterfactuals, we replace Blackrock with two different management companies with a cost structure analogous to Charles Schwab, which we estimate to be less efficient. The resulting welfare loss is slightly smaller than before and amounts to a 20% reduction in household welfare. Overall, our counterfactual exercises suggest that restricting the largest management companies to favor competition might ultimately hurt investors' welfare if those companies are also the most efficient ones.

Finally, in Proposition 8, we provide a closed-form expression of the equilibrium asset price multiplier with respect to investors' wealth. Our estimates imply that a 1% increase in household wealth increases the valuation of the equity index by 5.5%, which is in line with what the recent asset pricing literature has found. In the context of our model, the inelasticity of asset demand is driven by the structure of the asset management industry. The presence of large and multiproduct management companies exacerbates the price impact of a demand or cost shock. In Section 3.6.4, we perform a comparative static exercise and show that a reduction in fund initiation costs pushes management companies to introduce more funds which in turn reduces fees and further increases household demand for the index. Under fixed supply, asset prices will need to be higher to clear the asset market. Overall, our model rationalizes the high price multiplier often found in the empirical asset pricing literature through the competitive dynamics of large, heterogeneous and multiproduct management companies.

The rest of the paper proceeds as follows. Section 3.2 reviews the literature. Section 3.3 describes the model. Section 3.4 proves uniqueness and existence of the symmetric steady-state. Section 3.6 calibrates and estimates the model. Section 3.7 concludes.

3.2 Related Literature

Our paper contributes to the broad literature that studies theoretically and empirically the industrial organization of the asset management industry. Dermine, Neven and Thisse (1991) is one of the first contributions that considers strategic competition between funds. In their model, funds face demand from mean-variance investors with heterogeneous risk-aversion and choose where to locate on the mean-variance frontier. The equilibrium displays an Hotelling like property in which only two types of funds are supplied, one fully invested in the risk-free and the other fully invested in the market. Nanda, Narayanan and Warther (2000) propose a model of the mutual fund industry in which investors' heterogeneity in terms of liquidity needs implies that different load fee structures arise in equilibrium. Hortaçsu and Syverson (2004) show that presence of investors' search costs play a crucial role in rationalizing why homogeneous S&P 500 index funds with different fees are supplied. In our model we abstract from investors' heterogeneity but allow for multi-product management firms with heterogeneous production technologies who compete in a two-layers Cournot oligopoly.

Within this literature, a few papers highlight the importance of management companies in shaping the market structure and the proliferation of products in the asset management industry. Massa (2003) argues that fund families are incentivized to offer a broad menu of funds because investors value the possibility to switch across different funds belonging to the same family at no cost. Khorana and Servaes (1999) empirically analyze the determinants of mutual fund starts and show that scale and scope economies are among the factors that induce fund families to launch new funds. More recently, Betermier, Schumacher and Shahrad (2022) provide empirical evidence that incumbent families set up a large number of new funds in order to deter entry. The role of fund families and product proliferation are also crucial elements in our dynamic model.[4] In each period fund families decide how many

4. The importance of multi-product management companies is not limited to shaping the market structure of the industry. Gaspar, Massa and Matos (2006) provide empirical evidence of how fund families transfer performance across member funds to maximize family profits. Bhattacharya, Lee and Pool (2013) show

new funds to introduce taking into account that operating more funds will generate scale economies next period but will increase competition and reduce profits of existing funds.

Our work also contributes to the recent asset pricing literature that highlights the role of institutional investors in determining asset prices movements. In a static mean-variance framework Petajisto (2009) shows how the presence of demand for asset management services is enough to generate downward sloping demand curves for stocks. Starting from a simple portfolio choice problem Koijen and Yogo (2019) develop an equilibrium asset pricing model in which portfolio weights are function of stock characteristics. The model is estimated to match investors holdings and used in several applications to highlight the role of institutions in determining asset market movements. More recently, Haddad, Huebner and Loualiche (2022) extended the Koijen and Yogo (2019)'s characteristics-based framework by allowing investors to compete between each other in setting their trading strategies. They show that this type of strategic competition reduces the competitiveness of the stock market leading to inelastic demand curves.[5] In our model equilibrium asset prices are also related to the competitive behavior of large multi-product institutional investors. In particular, we show how technology primitives such us scale economies in the production of asset management services impact equilibrium asset prices.

Recently, the importance of large institutional investors has been also shown to be relevant for the efficiency of asset prices. In a recent paper, Kacperczyk, Nosal and Sundaresan

that large families offer mutual funds that only invest in other funds in the family and how these type of funds provide insurance against liquidity shocks. Berk, Binsbergen and Liu (2017) argue that fund families exploit their private information about their managers skill and create value by reallocating capital efficiently among managers. Ørpetveit (2021) shows empirically that management companies improve the quality of their existing funds in response to higher competition.

5. Many other papers have included institutional investors in asset pricing models among which He and Krishnamurthy (2013) and Basak and Pavlova (2013). The former paper develops a dynamic model in which institutional investors are constrained in their portfolio choice and study the impact on risk premia in bad times. The latter studies the asset pricing effects of delegated portfolio managers who care about their performance relative to a benchmark in a dynamic economy. More recently, Pavlova and Sikorskaya (2022) develop an empirical measure of benchmarking intensity and provide evidence of inelastic demand of active managers for stocks in their benchmarks.

(2022) consider an asset market with an oligopoly of large investors of exogenous sizes and study how market concentration affects price informativeness.[6] Although we abstract from the role price informativeness, our model endogenizes flows and market concentration leaving heterogeneity in production technologies to be the fundamental model primitive.

Finally, motivated by the increasing regulatory scrutiny toward the growth of index investing,[7] Schmalz and Zame (2023) propose a static equilibrium model with heterogeneous investors and show that the presence of an index fund might hurt investors' welfare if one takes into account the general equilibrium effect on asset prices. When the index fund enters the market or lowers its fee investors increase their stock holdings relative to their bond holdings, which leads to higher asset prices and, in turn, to lower asset returns. Although our model is different in several respects, we also look at how investors' welfare changes when the structure of the asset management industry changes, while taking into account the effects on asset prices. Our counterfactual analysis in section (3.6.3) suggests that restricting the largest passive asset managers to favor competition might reduce investors' surplus. This happens because the largest management companies are far more efficient than the rest of market and thus the efficiency loss that results from removing them hurts investors despite asset prices increase.

6. With the rise of passive investing the literature that studies how the presence of large passive investors affects the information embedded in asset prices is growing. See for instance: Bai, Philippon and Savov (2016), Baruch and Zhang (2022), Bond and Garcia (2022), Farboodi, Matray, Veldkamp and Venkateswaran (2021), Coles, Heath and Ringgenberg (2022), Malikov (2021) and Sammon (2022).

7. Regulators and antitrust legal scholars are investigating the consequences of the rise of passive investing on various economic outcomes. The trigger of many of the regulatory concerns has been a recent and growing literature that studies the anticompetitive effects of common ownership (Azar, Schmalz and Tecu (2018)), Posner, Scott Morton and Weyl (2017), Anton, Ederer, Gine and Schmalz (2017), Azar and Vives (2021), Backus, Conlon and Sinkinson (2021)). This literature asks whether product firms that share common owners, which in most cases are large passive asset managers, have less incentive to compete.

3.3 Model

Time is discrete and indexed by $t \in \{1, 2, ...\}$. We consider an economy populated by three types of agents: a representative household , mutual funds and management companies. The representative household allocates wealth between the mutual fund sector and a risk-free asset to finance consumption and takes expected return, variance and fees as given. The mutual fund sector is populated by a discrete number of identical funds that internalize household demand and that optimally choose their size to maximize dollar revenues. Each fund invests in the same underlying index and takes the total number of operational funds as given. Finally, each management company is responsible for fund initiation. Specifically, at each time t, each management company controls a number of pre-existing funds that carries from previous periods and chooses the number of new funds to create. We close the model and derive equilibrium market prices by assuming that mutual funds have a strict mandate to invest in the underlying index and that the index is available in fixed supply. The major contribution of the model is to describe the competitive dynamics of the mutual fund industry assuming that not only mutual funds but also management companies simultaneously and dynamically compete a la Cournot. Despite the complications created by the two layers of Cournot competition, we provide sufficient conditions under which the model admits a unique steady-state equilibrium.

We now proceed to describe in details the problem solved by each agent in the model.

3.3.1 Household

In each period, a representative household with log utility over consumption decides how much of its current wealth A_t to consume and how much to invest in the financial market. The investment opportunity set consists of two broad asset classes, namely a risk-free asset with return normalized to zero and a mutual fund sector. The mutual fund sector is populated

by a discrete number of identical funds that invest in the same underlying index and that charge fee f_t at time t.[8] Because all mutual funds are identical, each household is indifferent between investing in any specific fund and it will only choose the fraction of wealth to invest in the aggregate mutual fund sector. The size of each individual fund will then be determined via Cournot competition.

We assume that the index tracked by each mutual fund pays a constant dollar dividend D and we denote by P_t the index price at time t. Next, we define the net of fee index return at time $t+1$ as

$$1 + R_{t+1} = \frac{P_{t+1} + D}{P_t} - f_t. \tag{3.1}$$

Our representative household knows D and f_t but is not able to foresee the equilibrium path of asset prices. In other words, the household is not able to anticipate the effect that actions of mutual funds and management companies have on equilibrium asset prices. Instead, he perceives the index log net returns to evolve as a Gaussian stationary process

$$r_{t+1} \equiv \log(1 + R_{t+1}) = \rho_t - f_t + \sigma_t \varepsilon_{t+1}$$

where $\varepsilon_{t+1} \sim \mathcal{N}(0,1)$.[9] Letting w_t denote the portfolio weight on the mutual fund sector, the problem solved by the representative household at time t is

$$V(A_t) = \max_{C_t, w_t} \mathbb{E}_t \left[\sum_{s=t}^{\infty} \beta^{s-t} \log(C_s) \right] \tag{3.2}$$

$$\text{s.t. } A_{t+1} = (1 + w_t R_{t+1})(A_t - C_t) \tag{3.3}$$

8. We will discuss this product homogeneity assumption in Section 3.3.5.

9. When simple net returns are sufficiently small $\log(1 + R_{t+1}) \approx R_{t+1} = \frac{P_{t+1}+D}{P_t} - f_t - 1$ so that ρ_t can be interpreted has the household subjective belief about the next period capital gain and dividend yield.

with associated Euler equation given by

$$1 = \mathbb{E}_t \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-1} (1 + R_{t+1}) \right] \tag{3.4}$$

Under standard arguments, which we detail in Appendix 3.8, the household optimally consumes $C_t = (1-\beta)A_t$ and invests βA_t in the financial market. Moreover, household optimal portfolio allocation is given by

$$w_t = \frac{\mu_t - f_t}{\sigma_t^2}. \tag{3.5}$$

where $\mu_t \equiv \rho_t + \sigma_t^2/2$.

3.3.2 Mutual Funds

In any period t, each mutual fund takes the total number of funds in the market n_t as given and chooses its market share to maximize dollar profits. Given the optimal market share chosen by each fund, the equilibrium fee at time t is pinned down by the demand of the representative household in (3.5). In other words, we are assuming that, at each time t, mutual funds compete simultaneously and repeatedly a la Cournot. Each mutual fund internalizes that a higher individual market share leads to a higher market share of the aggregate mutual fund industry and to a lower fee that the household is willing to pay.

Let w_{it} denote the weight on mutual fund i in the portfolio of the representative household, and consider to rewrite household demand in (3.5) as

$$f_t = \mu_t - w_t \sigma_t^2 \tag{3.6}$$

Then fund i at time t solves

$$\max_{w_{it}} \; f_t w_{it} \beta A_t$$

subject to

$$f_t = \mu_t - w_t \sigma_t^2$$
$$w_t = \sum_{i=1}^{n_t} w_{it}$$

Taking the first order condition with respect to w_{it} we obtain fund i best response

$$w_{it} = \frac{\mu_t}{\sigma_t^2} - w_t. \tag{3.7}$$

Summing across funds yields the Cournot total quantity

$$w_t = \frac{\mu_t n_t}{\sigma_t^2 (n_t + 1)}, \tag{3.8}$$

and by replacing (3.8) in (3.6) we recover the equilibrium fee

$$f_t = \frac{\mu_t}{n_t + 1}. \tag{3.9}$$

Finally, we solve for the symmetric equilibrium w_{it} by replacing (3.8) in (3.7)

$$w_{it} = \frac{\mu_t}{\sigma_t^2 (n_t + 1)}. \tag{3.10}$$

In equilibrium at time t and conditional on n_t, each mutual fund i realizes dollar profits

$$\pi_t(n_t) = f_t w_{it} \beta A_t = \frac{\mu_t^2}{\sigma_t^2 (n_t + 1)^2} \beta A_t.$$

To save notation, we will rewrite dollar profits gained by each fund as

$$\pi_t(n_t) = \frac{\pi_t}{(n_t+1)^2} \tag{3.11}$$

where $\pi_t \equiv \frac{\mu_t^2}{\sigma_t^2}\beta A_t$.

3.3.3 Management Companies

Consider an oligopoly of M multi-product management companies indexed by $j \in \{1, 2, \ldots, M\}$. Each management company j enters time t with n_{jt-1} pre-existing funds and chooses the number of funds n_{jt} to operate in the current period with the objective of maximizing the present discounted value of dollar profits. Equivalently, the decision of management company j to operate n_{jt} funds at time t requires the creation or deletion of $n'_{jt} = n_{jt} - n_{jt-1}$ funds.

While pre-existing funds do not carry any cost for the controlling management company, opening a new fund is costly. We allow the initiation cost to depend on the size of management companies and we parameterize the cost of creating a new fund for management company j at time t as

$$C_j\left(n_{jt}, n_{jt-1}; c_j, \delta_j\right) = c_j n'_{jt} + \delta_j \left(\frac{n'_{jt}}{n_{jt-1}}\right)^2 n_{jt-1} \tag{3.12}$$

where $c_j > 0$ is the linear component of the initiation cost and δ_j captures an additional cost of adjusting the menu of funds, which we assume decreasing in the size of the management company, i.e. in the number of pre-existing funds n_{jt-1}. The suggested functional form for $C_j(\cdot)$ implies that management companies with a higher number of pre-existing funds face a lower initiation cost and it is motivated by the newly documented evidence that largest management companies are responsible for most of fund creation. We interpret this evidence

as suggesting that largest management companies face a lower cost of initiating a new fund and we incorporate this empirical fact in the model. Holding n_{jt-1} constant, the parameter δ_j captures how costly is for company j to adjust its menu of funds. A lower δ_j and a higher n_{jt-1} both reduce j's cost of launching a new fund.

What is the trade-off that management companies face when initiating a new fund? First, in the current period, there is an ambiguous effect on profits. On one hand, profits increase because the company operates one additional fund thereby increasing its market share. On the other hand, creating one additional fund increases competition in the mutual fund sector, leading to a decrease in fee f_t and profit $\pi_t(n_t)$. Second, expanding the current menu of funds carries the additional benefit of reducing the initiation cost in future periods.

Overall, management company j at time t solves the following dynamic problem

$$V\left(n_{jt-1}\right) = \max_{n_{jt}} n_{jt}\frac{\pi_t}{(n_t+1)^2} - C_j\left(n_{jt}, n_{jt-1}; c_j, \delta_j\right) + \beta V\left(n_{jt}\right) \tag{3.13}$$

subject to

$$n'_{jt} = n_{jt} - n_{jt-1}$$
$$n_t = \sum_{j=1}^{M} n_{jt}$$

Effectively, we are considering a dynamic game in which management companies compete simultaneously a la Cournot. For each management company j, the optimal strategies $n_{-jt} = (n_{j't})_{j' \neq j}$ chosen by the other enter management companies $j' \neq j$ enter the problem only through the total number of funds $n_t = n_{jt} + \sum_{j' \neq j} n_{j't}$.

3.3.4 Financial Market

The last aspect of the model that still has to be addressed is how the price of the index in which mutual funds are invested will be pinned down in equilibrium. To this end, we assume that mutual funds have a strict mandate to invest in the underlying index and that the index is available in fixed supply $\bar{Q}$. Letting Q_{it} denote the number of index shares demanded by mutual fund i at time t, then the assumption of strict mandate requires

$$Q_{it}P_t = w_{it}\beta A_t \qquad \forall i, t \tag{3.14}$$

Equation (3.14) simply states that, if mutual fund i has a strict mandate to invest in the index, then, at any time t, the dollar investment in the index (left hand-side) has to equal the total assets under management of mutual fund i (right hand-side). Summing (3.14) across funds and imposing market clearing yields

$$P_t = \frac{w_t(\beta A_t)}{\bar{Q}} = \frac{\mu_t n_t}{\sigma_t^2(n_t+1)}\frac{\beta A_t}{\bar{Q}} \tag{3.15}$$

where in the second equality we used equation (3.8). In equilibrium, the wealth A_t of the representative household will evolve according to the following law of motion:

$$\begin{aligned} A_{t+1} &= \beta A_t \left[1 + w_t\left(\frac{P_{t+1}+D}{P_t} - f_t - 1\right)\right] && (3.16)\\ &= \beta A_t \left[1 + \left(\frac{\mu_t n_t}{\sigma_t^2(n_t+1)}\right)\left(\frac{P_{t+1}+D}{P_t} - f_t - 1\right)\right] && (3.17)\\ &= \beta A_t + \bar{Q}\left(D + \Delta P_{t+1} - f_t P_t\right) && (3.18) \end{aligned}$$

where the second equality substitutes for the equilibrium portfolio weight w_t in (3.8) and the third equality uses expression (3.15). We stress that, because the household is not able to internalize the effect on asset prices coming from the actions of mutual funds and

management companies, then the law of motion of wealth derived in (3.18) will not in general be equivalent to the budget constraint used in (3.3).

3.3.5 Discussion of model assumptions

Before turning to the definition of equilibrium and proving existence and uniqueness of a steady state, we now discuss in detail some of our modelling assumptions. Although in some cases restrictive, all of the assumptions are needed to balance the model tractability and its ability to capture what we believe are the most relevant dynamics of the industry.

Myopic portfolio choice. In our model the optimal portfolio choice is myopic because our representative household has log preferences over consumption. Unless the belief process is independent and identically distributed over time, relaxing this assumption would affect our household portfolio choice in a way that would prevent us from obtaining a closed form solution for the portfolio weight w_t. In the case in which the belief process is time-varying the optimal portfolio choice would need to account for the incentives to hedge intertemporally. The resulting asset demand function would only be defined implicitly, making it hard to set up the supply side oligopolistic game in a tractable way. Overall, although the incentives to hedge intertemporally are important, our static demand framework allows us to derive a simple demand for asset management services in each period and to enrich the dynamics on the supply side without losing tractability.

Product homogeneity. In our model all funds are identical and in equilibrium will charge the same fee and manage the same amount of AUM. In practice though the funds offered by a management company are never perfectly identical. Even within the same investment category, funds in a company's menu might have slightly different holdings, different managers, different fee structures, tax benefits and so on. Allowing for this type of product heterogeneity would require extending the model in two directions: on the supply side, we would need to introduce some dimension of horizontal differentiation and characterize a fund with

a vector of both portfolio (e.g., type of holdings, factor exposures, etc.) and non-portfolio (e.g., management tenure, advertising, fund age etc.) characteristics. On the demand side, we would need to modify households' preferences for all these product characteristics to be valuable.

Product differentiation is without any doubt an important dimension through which investment firms compete to attract investors with heterogeneous preferences.[10] However, it should also be clear that introducing horizontal differentiation in our equilibrium framework would make it intractable and that is why we abstract from it. To further back up our homogeneity assumption, Tables 3.7 and 3.8, present some characteristics of the top 30 passive funds supplied in the Large Cap and Mid Cap sectors in 2018. The exposures to the 4 Carhart factors, the alphas and the gross-returns are similar across all funds especially within but also across the two sectors. Also, note that our model can be easily extended to include multiple index sectors with sector specific investors that do not substitute across sectors.[11]

Overall, we believe that this homogeneity assumption could be a good modelling compromise that would still allow us to study the competitive and asset pricing implications of fund proliferation while keeping the model tractable.

Investor learning. A large body of the literature on mutual funds has studied the so called flow-performance relationship.[12] According to the literature, past performance attracts new

10. Kostovetsky and Warner (2020) develop a textual measure of product differentiation and show that more differentiated/unique funds are able to attract higher inflows at least the first few years upon introduction. Abis and Lines (2022) use a k-mean mean clustering algorithm based on a textual analysis of fund prospectuses and show that funds are differentiated in groups and that investors withdraw money if funds tend to diverge from their prospectus strategy. Ben-David, Franzoni, Kim and Moussawi (2022) provide evidence that specialized ETFs that track niche portfolios are supplied to cater investors heterogeneous beliefs.

11. The case in which investors substitute across sectors, say because of diversification motives, would substantially complicate the oligopolistic game between funds and management companies. The reason is that funds are now also competing across sectors with products that have different characteristics.

12. See for instance the two seminal contributions Chevalier and Ellison (1997) and Sirri and Tufano (1998).

inflows regardless of whether performance persists or not. Building on this empirical findings, theoretical models studying the flow-performance relationship typically feature a learning component in which investors learn about unobserved managerial skills from past performance.[13] In our model we do not have investor learning because we are focusing on passive investment vehicles that track an underlying index. Therefore, we decided to not include investor learning in our dynamic model.

3.4 Equilibrium

3.4.1 Equilibrium definition

We are now ready to define the equilibrium of our dynamic game. Following the industrial organization literature on dynamic oligopolies, we restrict our attention to Markovian strategies i.e., strategies that are function of payoff relevant state variables.[14] From problem (3.13), we can see that the payoff relevant state-variables for management company j are its menu of funds active in the previous period n_{jt-1} as well as competitors' menu of funds active in the previous period $(n_{j't-1})_{j'\neq j}$. Indeed, management companies $j' \neq j$ choose an optimal strategy $(n_{j't})_{æ'\neq j}$ which is a function of $(n_{j't-1})_{j'\neq j}$. Because $(n_{j't})_{æ'\neq j}$ enter company j's problem through n_t, then the optimal strategy of each management company depends on its own menu of pre-existing funds as well as the menu of pre-existing funds of all its competitors. To preserve computational tractability, for each management company j, we restrict attention to strategies that are function only of company j's own state (in our

13. The seminal contribution here is Berk and Green (2004) which rationalizes the flow-performance relationship in a model with rational investors who learn about managers' alphas. More recently, Roussanov, Ruan and Wei (2021) extends the Berk and Green (2004) to allow for search friction as in Hortaçsu and Syverson (2004).

14. See for instance, Maskin and Tirole (1988), Ericson and Pakes (1995) and for a self-contained review Aguirregabiria, Collard-Wexler and Ryan (2021).

case, n_{jt-1}) and denote the policy function by $\alpha_j : [0,\infty) \to [0,\infty)$.[15]

Definition 2. *An equilibrium of our dynamic model consists in a profile of strategies* $\alpha^* = (\alpha_j^*)_{j=1}^M$ *with* $\alpha_j^* : [0,\infty) \to [0,\infty)$*, a path of asset prices* $(P_t)_{t=1}^\infty$ *and wealth* $(A_t)_{t=1}^\infty$ *such that:*

(1) in any period t*,* $\alpha^* = (\alpha_j^*)_{j=1}^M$ *is a pure strategy Markov perfect equilibrium such that for all* j

$$\alpha_j^*(n_{jt-1}) = \arg\max_{n_{jt}} \left\{ n_{jt} \frac{\pi_t}{(1+n_t)^2} - c_j(n_{jt} - n_{jt-1}) - \delta_j \left(\frac{n_{jt} - n_{jt-1}}{n_{jt-1}} \right)^2 n_{jt-1} + \beta V(n_{jt}) \right\}$$

where n_{jt} *denotes the number of company* j*'s funds active in the current period,* n_{jt-1} *the the number of company* j*'s funds active in the previous period,* $n_t = n_{jt} + \sum_{j' \neq j} n_{j't}$ *with* $n_{j't} = \alpha_{j'}^*(n_{j't-1})$ *and* $\pi_t = \frac{\mu_t^2}{\sigma^2} \beta A_t$*.*

(2) in any period t *the asset market clears,*

$$P_t = \frac{\mu_t n_t}{\sigma_t^2 (1+n_t)} \frac{\beta A_t}{\bar{Q}},$$

and the path of wealth solves,

$$A_{t+1} = \beta A_t + \bar{Q} \left(D + \Delta P_{t+1} - f_t P_t \right).$$

Before discussing existence and uniqueness of our equilibrium a few remarks are in order. First, our restricted Markovian strategies allow us to treat the Bellman of each of the M management companies as an independent single-agent dynamic problem. In other words, the strategy of each player does not depend on other players' states and thus, we can compute the envelope condition as in the standard single-agent frameworks. Second, we will assume

15. Weintraub, Benkard and Van Roy (2008) show that this restriction is appropriate in oligopolies with many firms and that as the number of firms increases the equilibrium converges to the unrestricted Markov perfect equilibrium.

that, when adjusting their menu of funds, management companies do not internalize the price impact generated by their actions. The profit earned by each management company indeed depends on expected returns and asset prices through the term π_t and asset prices in turn depend on the total number of funds n_t. Nonetheless we assume throughout that management companies take π_t as given.

3.4.2 Steady state definition and existence

While the computational complexity of the general model will require a numerical solution, we are able to formally characterize a steady-state equilibrium of our model characterized by

- $n_{j,t} = n_j > 0$ for any management company j and time t;
- $P_t = P > 0$ for any time t;
- $A_t = A > 0$ for any time t.

In other words, we are able to formally characterize a steady-state with constant index price, constant household wealth, and in which the dynamic game between management companies resolves with each company having incentive to keep the same number of funds over time.

We now turn to provide sufficient conditions for the existence and uniqueness of such a steady state equilibrium. We start by assuming that household subjective beliefs are constant over time, that is $\mu_t = \mu$ and $\sigma_t^2 = \sigma$ for all t. We maintain this assumption from here throughout the paper. It follows that, in steady state, the term π_t in (3.13) is also constant over time and equal to

$$\pi_t \equiv \pi = \frac{\mu^2}{\sigma^2}\beta A.$$

Proposition (7) provides sufficient conditions under which such steady state exists and is unique.

Proposition 7. *Let* $\tilde{\pi} \equiv \frac{D}{1-\beta}\frac{\beta\mu^2}{\sigma^2}$*, assume* $M\tilde{\pi} > (1-\beta)c$ *with* $c = \sum_j c_j$ *and, without loss of generality, let* $\bar{Q} = 1$*. Then, there exists a unique steady-state* $\{(n_j)_{j=1}^M, P, A\}$ *such that:*

(1) for any management company j *and any period* t*,* $n_{jt} = n_j = \alpha_j^*(n_j)$ *satisfies*

$$n_j = \frac{1+n}{2} - \frac{(1-\beta)}{2\pi}c_j(1+n)^3 \tag{3.19}$$

where $n = \sum_{j=1}^M n_j$ *and* $\pi = \frac{\mu^2}{\sigma^2}\beta A$*;*

(2) the market clearing price $P_t = P$*, the wealth* $A_t = A$ *and the total number of funds* n *solve simultaneously*

$$A = \left(\frac{1}{1+\zeta(n)}\right)\frac{D}{1-\beta} \tag{3.20}$$

$$P = \frac{\mu}{\sigma^2}\frac{n}{1+n}\left(\frac{1}{1+\zeta(n)}\right)\frac{\beta}{1-\beta}D \tag{3.21}$$

$$\tilde{\pi}(M + n(M-2)) = (1-\beta)c(1+n)^3(1+\zeta(n)) \tag{3.22}$$

where

$$\zeta(n) \equiv \frac{\mu^2}{\sigma^2}\frac{\beta}{1-\beta}\frac{n}{(1+n)^2}. \tag{3.23}$$

Moreover, $n_j > 0$ *and company* j *remains active if* $\frac{\pi}{(1+n)^2} > (1-\beta)c_j$*.*

Proof: See Appendix 3.8.

Equations (3.20) and (3.21) describe the equilibrium wealth and asset prices as functions of the equilibrium number of funds. Equation (3.20) suggests that the steady-state wealth A is proportional to the present discounted value of future dollar dividend D where the constant of proportionality depends on n, i.e. on the competitive outcome among management

companies. In particular, it is easy to notice that $\zeta(n) > 0$ and $\zeta'(n) < 0$ for $n > 1$. Thus, when competitive forces push companies to initiate a higher number of funds n, then $\zeta(n)$ declines and the steady-state wealth A increases. In the limit for $n \to \infty$, then $\zeta(n) = 0$ and $A = D/(1-\beta)$, i.e. the steady state wealth converges to the present discounted value of the dollar dividend D.

Similarly, according to equation (3.21), higher steady-state n leads to a higher index price P. In the limit for $n \to \infty$, we now have $P = \frac{\mu}{\sigma^2}\frac{\beta}{1-\beta}D$. More generally, equation (3.21) relates the equilibrium index price to the marginal cost of initiating new funds and thus microfounds the price impact of institutional investors in terms of the technological primitives of the asset management industry. In section 3.6.4, we will use equations (3.20), (3.21) and (3.22) to characterize the steady-state index price multiplier with respect to household wealth and we will show that this suggested measure of price impact crucially depends on the competitive outcome in the mutual fund sector. We will further perform a comparative static exercise to explore how the steady-state equilibrium, including the suggested measure of elasticity, vary with the dollar dividend D and the total fund initiation cost c.

While the result in Proposition (7) guarantees existence and uniqueness of a steady state in which all companies have no incentives to create additional funds and the index asset price is constant, we know less about the path $\{(n_{jt})_{j=1}^{M}, P_t\}_{t=1}^{T}$ that leads to such steady state. In the next section we provide a numerical algorithm that, for a given initial condition on the number of active funds $(n_{j0})_{j=1}^{M}$ and a given terminal date T, finds the optimal path if such path exists. The algorithm can be used to solve the model numerically and thus to derive the optimal path that, for given initial conditions, leads to the steady-state characterized in this section. For the purpose of this paper, we will use the algorithm to solve the model numerically and estimate the unobserved parameters in the cost function of management companies.

3.4.3 Numerical solution for the equilibrium path

The complexity of the problem prevents us from deriving formal properties of the equilibrium path out of the steady state. While formalizing its existence, convergence and stability properties is beyond the scope of the paper, the goal of our model is still quantitative and, as we will see in the next sections, we will estimate the model using data from the US mutual fund industry. With such goal in mind, in this section we propose a numerical procedure that, given proper initial and terminal conditions, allows to derive the equilibrium path if such path exists.

Our algorithm amounts to solving two fixed points, one nested into the other, that for a given set of initial conditions and parameter values, finds the optimal path of fund initiation, index price and household wealth. The numerical procedure can be summarized in the following steps:

Step 0. Set exogenous parameters to be kept constant throughout the algorithm:

- Fix exogenous parameters $\left\{\sigma,\ D,\ \mu,\ M,\ (c_j)_{j=1}^{M},\ (\delta_j)_{j=1}^{M},\ \bar{Q},\ \beta\right\}$.
- Fix the initial household wealth A_0.
- Fix the initial number of funds managed by each company j, $(n_{j0})_{j=1}^{M}$.
- Fix a terminal date T and the terminal number of funds managed by each company j, $(n_{jT})_{j=1}^{M}$.

Step 1. Solve the inner loop for a given path of asset prices $(P_t)_{t=1}^{T}$ and household wealth $(A_t)_{t=1}^{T}$ as follows:

- Construct the path for $(\pi_t)_{t=1}^{T}$.
- Guess a path for the number of funds managed by each company j: $\left(\left(n_{jt}^{(k)}\right)_{t=1}^{T}\right)_{j=1}^{M}$.

- Use euler equation (3.35) to find a new path $\left(\left(\tilde{n}_{jt}^{(k)}\right)_{t=1}^{T}\right)_{j=1}^{M}$.

- Update the path of funds using

$$n_{jt}^{(k+1)} = n_{jt}^{(k)} + \chi_n \left(\tilde{n}_{jt}^{(k)} - n_{jt}^{(k)}\right) \quad \forall j, t. \tag{3.24}$$

- Repeat until convergence.

Step 2. Run the outer loop to find the equilibrium path of index price and household wealth:

- Guess a path of prices $\left(P_t^{(q)}\right)_{t=1}^{T}$ and household wealth $\left(A_t^{(q)}\right)_{t=1}^{T}$.

- Run inner loop as in Step 1 to obtain $\left(\left(n_{jt}^{(q)}\right)_{t=1}^{T}\right)_{j=1}^{M}$.

- Use market clearing in (3.15) and the law of motion of wealth in (3.18) to find new paths $\left(\tilde{P}_t^{(q)}\right)_{t=1}^{T}$ and $\left(\tilde{A}_t^{(q)}\right)_{t=1}^{T}$.

- Update price and wealth using

$$P_t^{(q+1)} = P_t^{(q)} + \chi_p \left(\tilde{P}_t^{(q)} - P_t^{(q)}\right) \tag{3.25}$$

$$A_t^{(q+1)} = A_t^{(q)} + \chi_a \left(\tilde{A}_t^{(q)} - A_t^{(q)}\right) \tag{3.26}$$

- Repeat until the maximum of $||P^{(q+1)} - P^{(q)}||_\infty$ and $||A^{(q+1)} - A^{(q)}||_\infty$ is below some tolerance

To sum up, for a given set of parameter values, the routine just described starts in Step 2 with a guess for the equilibrium index price and wealth. It then moves to the inner loop (Step 1) and solves for the equilibrium number of funds taking the path of index price and wealth as given. Finally, it returns to Step 2 to update the equilibrium price and wealth. This routine is repeated until convergence. It is a nested procedure because the fixed point

that solves for the Markov perfect Nash equilibrium is solved within each iteration of the fixed point that solves for the market clearing price and wealth evolution.

3.5 Data

Before turning to the estimation of our model we overview our data sources and how we constructed our estimation dataset.

3.5.1 Data sources

We obtained data on US mutual funds from the Center for Research in Security Prices (CRSP) which we accessed through the Wharton Research Data Services (WRDS). The data provide detailed information on US mutual funds at monthly frequency starting from 1961 but we restrict the sample from year 2000 to 2020 for the reasons we describe in the following subsection.

The data is at the share class level but we collapse everything at the fund-by-year level. Moreover, we focus on US domestic equity funds that, according to the CRSP investment objective classification, belong to the Large Cap, Mid Cap and Small Cap sectors. Among those, we identify passive funds as either index funds or ETFs as classified by CRSP. The resulting sample contains around 16,500 fund-by-year observations of which around 3,700 are passive investment vehicles.

Table (3.5) presents some summary statistics of our data. The average amount of asset under management at the end of year is around 2 billions but the distribution is quite skewed due to the presence of extremely large funds. The average monthly gross return in a given year is around 0.9%, with an average monthly alpha of 0.04% and an average market beta of 0.97. These latter are estimated for each year and each fund from a monthly regression of gross returns on the 3 Fama-French factors plus momentum including observations from

the previous 3 years. Finally, the average market share at the management company level is around 1.7%, although also in this case the distribution is very skewed because, for most years, more than 50% of the market is captured by the five biggest management companies.

Table (3.6) replicates table (3.5) restricting the sample to passive funds only. As expected passive funds tend to be cheaper with an average expense ratio of 0.5% and larger, managing an average of 5.7 billions of assets. On average passive funds also seem to deploy more funds with an average of 4.5 funds per management company.

3.5.2 Data construction

We now discuss in detail the way we constructed our final dataset which we will use for estimating the model in the next section.

Filling missing of fund and company identifiers. Information about US mutual funds collected by CRSP is provided at the share class level. Data on returns and asset under management are at the monthly frequency whereas information on fund characteristics are provided quarterly. The first thing we do is to aggregate all share classes of the same fund in one single observation so that the resulting dataset is at the fund level. To this end, we exploit a grouping variable constructed by CRSP (denoted by *crsp_cl_grp*) that contains an unique code for all share classes that belong to the same fund. This variable is available starting from 1999 which is the main reason for why we restrict our sample to start from 2000. To identify funds of the same share class when *crsp_cl_grp* is not available we rely on the WFICN identifiers and on fund names. Fund names in CRSP are useful because contain three types of information: the name of the management company, followed by the name of the fund, followed by the type of share class. The former two are separated by a colon while the latter by a semicolon. Following this rule we parse each fund name in each month in three parts and then assign the same *crsp_cl_grp* to funds with the same fund name (i.e., the same second part of the name) in the same quarter. This procedure leaves us with 625

share class by quarter observations with a missing *crsp_cl_grp* out of more than 2 millions share class by quarter observations.

Key to our analysis is the role of management companies as fund initiators. In the data, we identify the management company that offers each fund using a unique management company identifier *mgmt_cd*, provided by CRSP, which is available starting from December 1999. Roughly 11% of share class by quarter observations have a missing *mgmt_cd* which we refill again exploiting the information available in the fund name. The first part of each fund name corresponds to the name of the management company; whenever missing we assign the same *mgmt_cd* to funds that feature the same management company name in the same quarter. This procedure fills around 60% of the missing *mgmt_cd*. Whenever this procedures fails because of mistakes in fund name spellings we refill *mgmt_cd* manually.[16] Overall, we were not able to identify the controlling management company for less than 1% of share class by quarter observations.

Aggregation of share classes and further cleaning. After the refilling procedure, we merge the quarterly level data on funds' characteristics (which include the *crsp_cl_grp* and *mgmt_cd* identifiers) with the monthly data on returns and AUM. Then, for each month we aggregate share classes of the same fund into one observation based on the *crsp_cl_grp* identifier. To do so we sum the end of month AUM of all share classes and take averages of other relevant variables, such as monthly returns and expense ratios, weighting by the AUM at the end of the previous month. Finally we only keep domestic equity funds and, to remove incubation bias, we drop funds that we observe for less than 12 months and whose

16. In some cases, mergers and acquisitions between companies create mismatches between fund names and the *mgmt_cd* code provided by CRSP which we manually correct whenever possible. For instance, after Blackrock acquired the iShare business from Barclays in June 2009 the *mgmt_cd* has not been updated accordingly. In this case there were two *mgmt_cd* codes "BZW" and "BLK" for Blackrock but we replaced "BZW" with "BLK" after 2009. Similarly, we replaced "PDR", the *mgmt_cd* for PDR services LLC owned by the American Stock Exchange, with "SSB" the *mgmt_cd* for State Street Bank which acquired the SPDR ETF license from PDR services LLC in 2005.

AUM are less than 15 millions.[17] The resulting dataset contains around 650,000 fund by month observations.

Dataset for model estimation. Our model focuses on homogenous passive investment vehicles that track an underlying index. In the data we identify passive funds using the variables *et_flag* and *index_fund_flag* and consider as passive both index funds and ETFs. Moreover, we restrict ourselves only to the Large Cap, Mid Cap and Small Cap sectors as identified by the *crsp_cl_grp* variable constructed by CRSP.[18] The reason is that more than half of pure index funds belong to these sectors and, as shown in Table 3.8 these products seem to be sufficiently homogeneous in terms of the risk-return profile they offer. Finally, we collapse everything at the year level and we obtain a dataset of 16,500 fund by year observations of which 3,700 are passive.

3.6 Model estimation

Using the numerical algorithm discussed in section 3.4, we now turn to estimate the model and discuss the results. Specifically, in section 3.6.1, we provide details of the estimation procedure. In section 3.6.2, we comment on the results, including the ability of the model to match targeted as well as untargeted moments. We then turn in section 3.6.3 to perform a series of counterfactuals devoted to assessing the contribution of each management companies to fund proliferation, fee and household surplus. Finally, section 3.6.4 shows that the model is suitable to speak to a growing literature that focuses on the asset pricing implications of inelastic demand for financial assets. Despite being completely untargeted, we estimate that a 1% increase in household wealth increases the steady-state valuation of the equity index

17. To identify domestic equity funds we exploit the variable *crsp_obj_cd* which classifies funds based on their investment style. The variable is constructed by CRSP building on Strategic Insights, Wiesenberger, and Lipper objective codes

18. Funds classified to belong to these sectors determine their holdings primarily on market capitalization considerations.

by 5.5%. The resulting 5.5 steady-state multiplier is not only aligned with what the previous literature has documented, but it also provides an alternative microfoundation of inelastic markets, namely the competition among oligopolistic investment management companies.

3.6.1 Estimation procedure

The estimation procedure relies on calibrating a subset of the parameters while inferring a second subset of parameters from data. Because our model abstracts from product differentiation, we estimate the model to match features of mutual funds classified as either Large Cap or Mid Cap in CRSP. In other words, we limit ourselves to passive funds that track reasonably mature firms and exclude instead mutual funds that track growing or developing companies. Table 3.1 summarizes the calibrated inputs.

Parameter	Description	Value
$\frac{D}{P_0}$	Dividend yield	2.14%
σ	Volatility	25.09%
μ	Expected return	6.12%
β	Discount factor	0.98
M	Number of management companies	6
A_0	Initial wealth	1.00
$\bar{Q}$	Supply of shares	1.00
T	Terminal date (years)	20

Table 3.1: Calibrated inputs

From our dataset, we estimate a dividend yield on the Russell 2000 equal to 2.14%. We then calibrate the dollar dividend D to match a dividend yield equal to 2.14% at time $t = 0$. We set household expected return μ to match the average return on the Russell 2000 index which we estimate equal to 6.12%. Similarly, we set the return volatility σ to match the standard deviation of the Russell 2000, equal to 25.09%. Since our attention is focused on

Large Cap and Mid Cap funds, we use the Russell 2000 rather than the S&P500 as the counterpart of the equity index in our model. We set the number of management companies equal to 6. This choice is motivated by the newly documented evidence that the top five management companies behave very differently compared to other management companies and are responsible for most of mutual fund proliferation. For this reason, we directly model competition among the top five firms and classify all other management companies in one residual group (from here on, we will refer to this residual group as the outside management company, indexed by $j = 0$). We identify the top management companies as the five firms with the highest average annual market share throughout our sample.[19] We further normalize both household initial wealth A_0 and the supply of index shares $\bar{Q}$ to one. Finally, we set the terminal date $T = 20$ to match the length of our dataset which ranges between 2000 and 2020. For the purpose of our solution algorithm, we then use as terminal condition $(n_{jT})_{j=0}^{5}$ the number of funds that we observe in our dataset for each management company in 2020.

While all the parameters discussed so far can be easily obtained from data or can be reasonably linked to observables, the same is not true for the parameters $(c_j)_{j=0}^{5}$ and $(\delta_j)_{j=0}^{5}$ that characterize the cost function of the management companies in our model. For this reason, we estimate both set of parameters directly from data using the following estimation procedure.

Let $\theta = (c_j, \delta_j)_{j=0}^{5}$ denote the set of parameters to be estimated. We estimate θ by solving

$$\min_{\theta} \sum_{s=1}^{S} \sum_{j=0}^{5} \left(\Lambda_{sj}(\theta) - \Lambda_{sj}\right)^2. \tag{3.27}$$

where $\Lambda_{sj}(\theta)$ denotes the s^{th} moment for management company j implied by the model and

19. The market share of a given management company in a given year is simply computed as the sum of net assets across all funds operated by the management company, rescaled by the sum of net assets across all funds that appear in our dataset in a given year.

expressed as a function of the unknown parameters in θ. On the other hand, we denote by Λ_{sj} the empirical analogue of $\Lambda_{sj}(\theta)$ observed in the data.

From the problem solved by a generic management company j in our model, it can be noticed that c_j governs the linear trend in its menu of funds. Thus, for given initial condition n_{j0} on the number of funds that management company j operates at time $t = 0$, c_j is tightly linked to the average number of funds that management company j originates in each period. Differently from c_j, δ_j governs how costly is for management company j to adjust and restructure its menu of funds. Specifically, given two management companies j and j' at time t with $n_{jt-1} = n_{j't-1}$, if $\delta_j < \delta_{j'}$ then adjusting the menu of funds is less costly for company j. In other words, management company j can more easily adapt its supply of funds without incurring in large adjustment costs. In mathematical terms, δ_j is directly related to the curvature across time in the equilibrium number of funds offered by company j, with lower δ_j translating into higher curvature in the equilibrium path.

Informed by the above discussion, we select the following two moments for a given management company j

$$\Lambda_{1j} = \sum_{t=1}^{T} \frac{\Delta n_{jt}}{T} \quad \forall j \tag{3.28}$$

$$\Lambda_{2j} = \sum_{t=1}^{T} \frac{\Delta\left(\Delta n_{jt}\right)}{T} \quad \forall j \tag{3.29}$$

In words, Λ_{1j} captures the average creation rate in absolute terms of management company j, which allows to pin down c_j. On the contrary, Λ_{2j} captures the concavity/convexity of management company j creation rate over time, and it allows to pin down δ_j. For each of the top five management companies, both moments are computed from the time-series of the number of funds operated by the management company between 2000 and 2020. For

the outside management company, both moments are computed from the time-series of the average number of funds operated by non-top five management companies over the same time interval. Finally, notice that the estimation problem involves $6 \times 2 = 12$ moments and $6 \times 2 = 12$ unknowns, so that it is exactly identified.

Concretely, we employ the following steps to obtain an estimate of θ:

- At the end of each iteration in the nested fixed loop described in section 3.4.3, we compute $\Lambda_{1j}(\theta)$ and $\Lambda_{2j}(\theta)$ for any management company j.

- Given Λ_{1j} and Λ_{2j} from data, we form the objective function in equation (3.27).

- We iterate over θ until the objective is minimized

The minimization works robustly and it takes around three seconds to solve one iteration on a standard portable computer.

We conclude this section by reporting in table 3.2 summary statistics for the six management companies used in our estimation procedure. In reporting the last row, we first construct the outside management company by averaging in each year across all non-top five management companies and by subsequently averaging in the time-series.[20]

The top five management companies have reported, on average, a cumulative market share of 85.76%. In other words, differentiating the top five management companies and regrouping all other firms allow modeling directly more than 80% of the market on average. A second relevant feature of the data is a clear positive relation between the average market share and the average number of controlled funds. This feature is consistent with the mechanisms in our model where, given the absence of fund differentiation, a management company can increase its market share only by increasing the number of funds it operates. The last three columns in table 3.2 further provide three set of parameters that directly

20. This practice has the shortcoming that average market shares do not generally sum to one, but it has the advantage that the outside management company can be interpreted as a representative "small" management company.

Management company	Share	Num. of funds	n_{j0}	n_{jT}	Λ_{1j}	Λ_{2j}
Vanguard	46.73%	6.95	4	9	0.25	0.00
State Street	17.25%	6.24	1	10	0.45	-0.05
Blackrock	9.89%	11.90	2	13	0.55	-0.36
Fidelity	9.61%	3.10	2	8	0.30	0.05
Charles Schwab	2.28%	2.62	2	4	0.10	0.00
Outside MC	0.17%	1.78	1.33	1.97	0.03	-0.01

Table 3.2: Summary statistics and estimated inputs

enter the estimation procedure. Columns (4) and (5) report the number of funds that each management company used to operate in year 2000 and 2020 respectively, which we employ as initial and terminal conditions in the model estimation. Columns (6) and (7) provide the empirical analogue of the moments we use to estimate the model. Blackrock and State Street are characterized by the highest absolute rate of fund creation Λ_{1j} but also by the most concave transition pattern Λ_{2j}. These features already point to the presence of both a low linear cost c_j and low adjustment cost δ_j. Both the rate of fund creation, as well as the concavity of the transition pattern, are lower for Vanguard, Fidelity and Charles Schwab, but well above the corresponding moments reported for the outside management company. Interestingly, Fidelity is the only management company with positive Λ_{2j}, determined by the fact that Fidelity started engaging in fund creation only in recent years, after 2015. These features of data are confirmed by figure 3.4, which reports the time-series of the number of funds controlled by the top 5 management companies.

3.6.2 Results

We use the procedure as well as the moments discussed in section 3.3 to estimate the model. We start by reporting in table 3.3 the estimated vector of parameters θ across the six management companies considered in the estimation. For a direct comparison between estimated

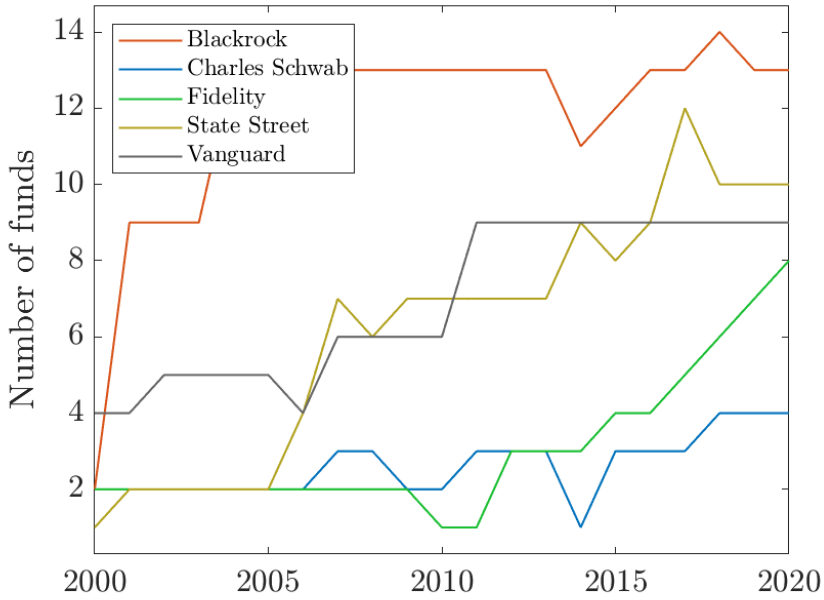

Figure 3.4: Number of funds operated by each of the top five management companies over time.

parameters and target moments, we reinclude n_{j0}, Λ_{1j} and Λ_{2j} in table 3.3 as well.

Management company	n_{j0}	n_{jT}	Λ_{1j}	Λ_{2j}	c_j	δ_j
Vanguard	4	9	0.25	0.00	0.0099	7.9003
State Street	1	10	0.45	-0.05	0.0007	3.0815
Blackrock	2	13	0.55	-0.36	0.0004	0.0001
Fidelity	2	8	0.30	0.05	0.0871	5.7127
Charles Schwab	2	4	0.10	0.00	0.0584	7.6989
Outside MC	1.33	1.97	0.03	-0.01	0.0238	5.0298

Table 3.3: Estimated parameters

With the lowest linear cost c_j and the highest speed of adjustment δ_j, Blackrock is the most efficient management company.[21] Such efficiency allowed Blackrock to increase

21. In the estimation we considered Blackrock and Barclays an unique entity even before their merger in 2009 when Blackrock acquired the Barclays' iShare business. Before the acquisition Blackrock market share

massively the number of controlled funds from 2 to 13 throughout the sample, with 0.55 funds created on average in each year. Compared to the beginning of our sample, Blackrock managed to become an absolute industry leader by 2020. The second most efficient firm is State Street. While in 2000 State Street was controlling only one fund (less than the average number of funds controlled by the outside management companies), it managed to create 0.45 funds per year on average, concluding the 2020 with 10 funds, second only to Blackrock. In 2000, Vanguard controlled more funds than any other management companies. The rate of fund creation, however, has been lower for Vanguard than for Blackrock and State Street. Finally, Fidelity and Charles Schwab appear as the least efficient firms among the top five management companies, although Fidelity experienced a significant bounce up that started in 2015.

We next turn to validate our estimated parameters. Using the estimated vector of parameters θ, we reconstruct the time-series of n_{jt} for each management company $j \in \{0, ..., 5\}$ as implied by the model solution. We further average n_{jt} across the top five management companies $j \in \{1, .., 5\}$ in each period. Figure 3.5 compares the time-series of n_{0t} and $\sum_{j=1}^{5} \frac{n_{jt}}{5}$ in model vs data.

The model is able to exactly match the high creation rate observed for the top management companies as well as the low rate of fund creation observed for other management companies. Despite our model is estimated from a set of exactly identified equations, we believe these estimates provide a useful quantitative benchmark to explain dynamics in this industry. This is confirmed by the fact that the estimated model is able to match extremely closely also untargeted moments and, in particular, the secular decline in average fee charged by Mid and Large Cap passive funds. Figure 3.6 compares the value-weighted fee observed

was small whereas Barclays was one of the top 5, the opposite happens after the acquisition. Another way to interpret this is thinking to iShare to be itself a multi-product company and according to our estimate the most efficient one. In practice, although the owner of the iShare business changed, iShare has always been one of the market leader since the early 2000.

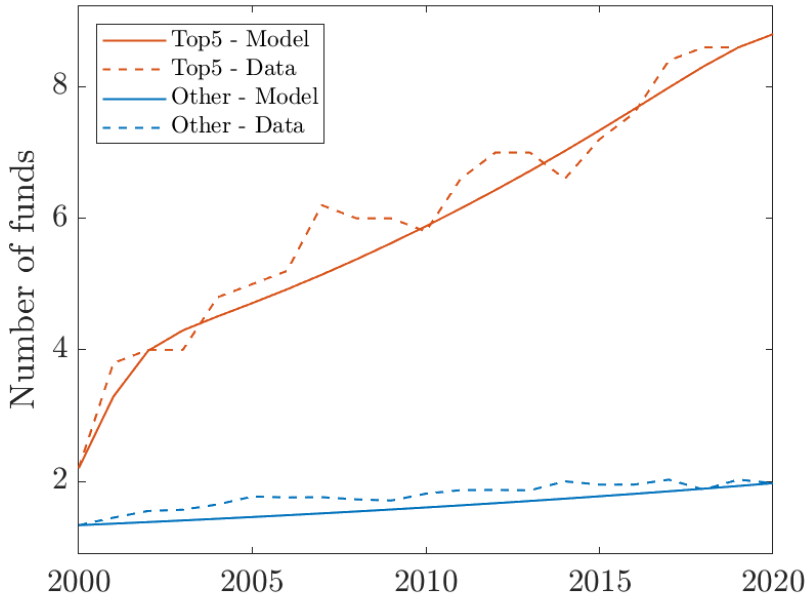

Figure 3.5: Time-series of the average number of funds operated by the top five management companies as well as the number of funds operated by the outside management company in model vs data. In data, the time-series of the number of funds operated by the outside management company is computed as simple average of the number of funds operated by all non-top five management companies

in data against the equilibrium fee implied by the model and estimated using equation (3.9).

Lastly, we compare the revenues gained by top five management companies with the revenues gained by the outside management company in the estimated model. We report the output in figure 3.7.

Management companies compete with each other over time and try to gain market share by creating new funds. As the number of funds in the market increase, the equilibrium fee declines and revenues decline as well both for top five management companies and for the outside management company. However, because the top five management companies are more efficient than the outside management company, they can efficiently use fund initiation as a tool to saturate the market, leading the outside management company to earn close to zero revenues by the end of the sample.

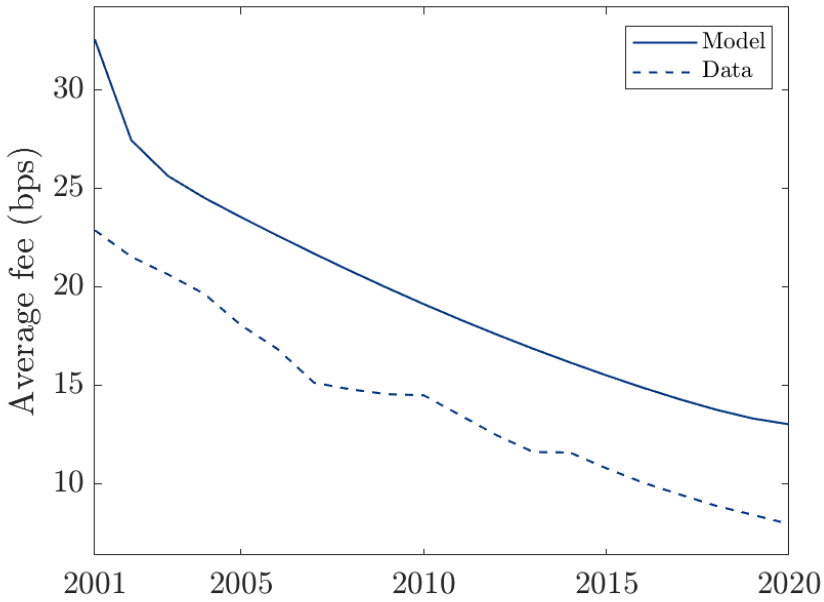

Figure 3.6: Time-series of value-weighted fee from data and equilibrium fee estimated from the model. The value-weighted fee from data is estimated, for each year, by averaging the expense ratio reported by CRSP for each fund with weights proportional to lagged total net assets.

3.6.3 Counterfactuals and welfare analysis

We now turn to the key section of the paper. Using the estimated model, we perform a series of counterfactuals devoted to understand the contribution of each management company to the secular decline in fees and to consumer surplus.

For each management company $j \in \{0, .., 5\}$ we start by fixing the initial number of funds n_{j0} at the level observed in 2000, the beginning of our sample. We further fix c_j and δ_j to the estimates obtained and discussed in section 3.6.2. Using the calibrated parameters in Table 3.1, we numerically solve for the model equilibrium over a long horizon which we set equal to $T = 100$ years. The model solution allows us to derive the equilibrium path for the number of funds n_{jt} held by each management company, the total number of funds n_t, the fee f_t and the revenues earned by each management company j. We further construct consumer surplus as

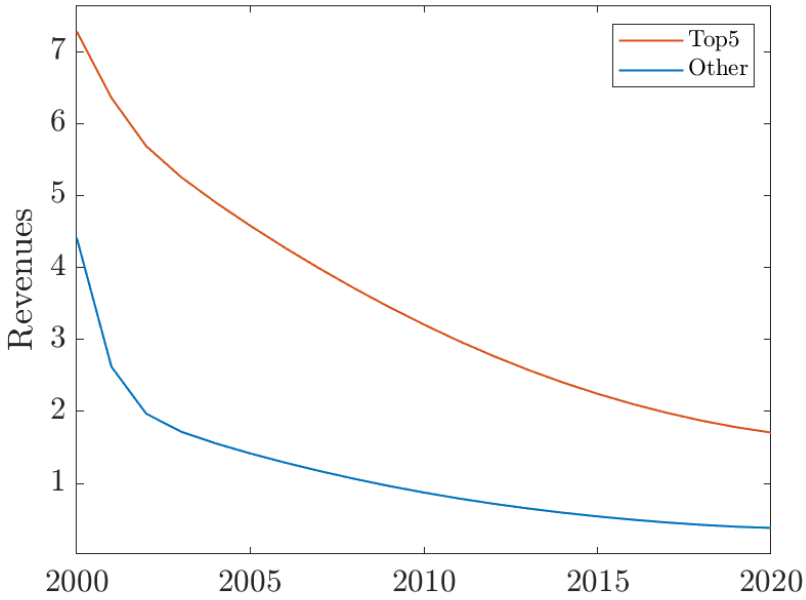

Figure 3.7: Estimated time-series of average revenues gained by the top five management companies and by the outside management company in the estimated model.

$$S = \sum_{t=0}^{99} \beta^t \log(C_t) + \beta^{100} \frac{\log(C_{100})}{1-\beta}$$

The equation for S implicitly assumes that household consumption remains constant at C_{100} for all $t \geq 100$. While this assumption is not an equilibrium outcome, setting the terminal date $T = 100$ so far away in the future implies that the terminal value $\beta^{100} \frac{\log(C_{100})}{1-\beta}$ only accounts for 2.74% of household overall surplus S.

Given the model solution, we perform a series of counterfactuals devoted to understanding the impact of each management company on equilibrium outcomes. Specifically, we solve the model after removing each management company j, one at a time. For each of the remaining management companies $j' \neq j$, we keep the same initial condition $n_{j'0}$ and the same estimates c'_j and δ'_j. This procedure allows us to construct the counterfactual equilibrium

path for number of funds, fee, revenue and consumer surplus that would have prevailed if management company j had not been operational.

Figure 3.8 provides the results of our counterfactual analysis. Each bar is labelled after the name of the management company that is excluded in the counterfactual of interest.

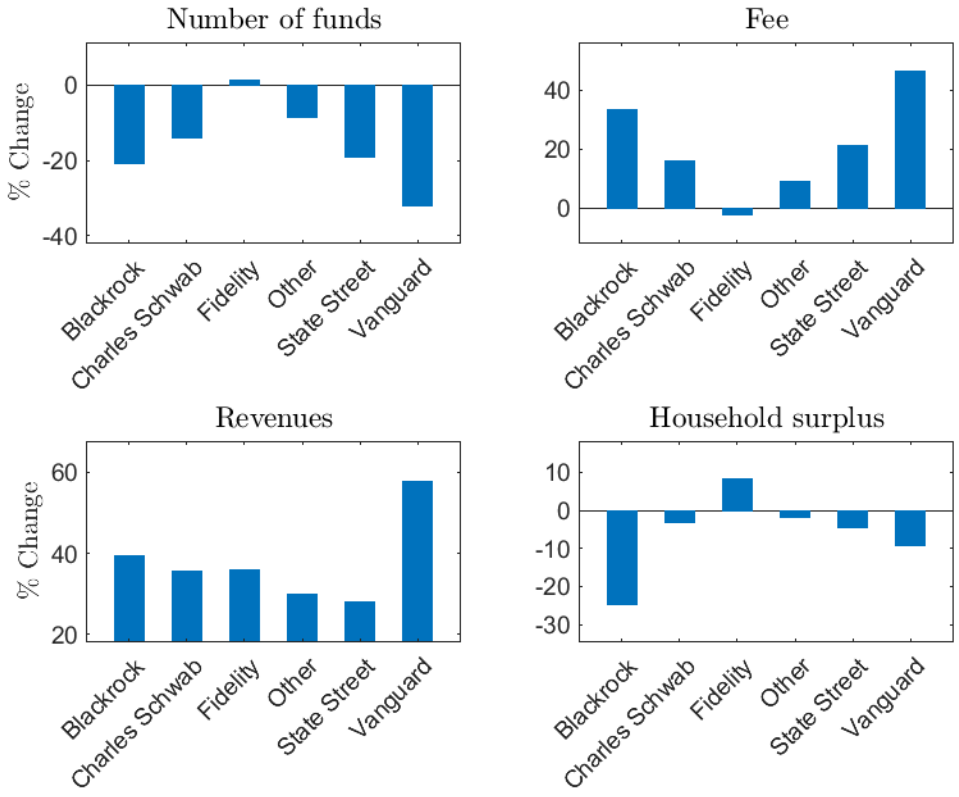

Figure 3.8: Percentage change in average number of funds, average fee, revenues and consumer surplus in each counterfactual compared to the model solution. For number of funds, we report the percentage change in average number of funds, where the average is computed across time. For the fee, we report the percentage change in the average fee, where the average is computed across time. For revenues, we report the percentage change in average revenues, where the average is computed across time and management companies.

We start from discussing the top-left panel, which reports the percentage change in average number of funds in each counterfactual compared to the model solution. Removing Vanguard would lead to the largest decline in the number of funds operating in the market and equal to 31.95%. Removing Blackrock and State Street would also lead to a signifi-

cant decline in the number of funds, equal to 21.00% and 19.20% respectively. Excluding Charles Schwab would decrease the number of operating funds by 14.10% while, interestingly, excluding Fidelity would leave the number of funds basically unaltered.

Turning to fees, the pattern is symmetric compared to the one seen for the number of funds. Removing Vanguard and Blackrock would significantly decrease competition in the mutual fund industry, leading to a 45.88% and 33.71% increase in the average fee respectively. The counterfactual increase in fee would be lower but still significant after removing State Street or Charles Schwab, equal to 21.32% and 16.12% respectively. The large increase in fees after removing Vanguard would significantly boost management companies revenues by 57.75%. The increase in average revenue would instead lie between 28.02% and 39.25% if any other management company is removed from the market.

Finally, we turn to household surplus. Removing Blackrock would lead to the largest decline in household surplus, equal to 24.64%. This is expected since Blackrock is by far the most efficient firm based on the estimates discussed in section 3.6.2. The second largest decline in household surplus is equal to 9.19% and is observed after removing Vanguard. Removing other management companies would lead to a change in household surplus between -4.57% and 8.02%.

Overall, the results so far suggest that large and efficient management companies play a crucial role in shaping the competition in the mutual fund sector. At the same time, the decline in household surplus may appear mechanical since removing any management company from the market would trivially lead to lower competition in the mutual fund sector. Lower competition would in turn increases equilibrium fees thereby reducing household welfare. We show that this is not the case. Specifically, we conduct a counterfactual where Blackrock is replaced by two additional management companies equal to Charles Schwab. This implies that the number of management companies competing in the industry increases from 6 to 7. We report in figure 3.9 the percentage change in the number of funds, fee, revenues and

household surplus between this counterfactual and the model solution.

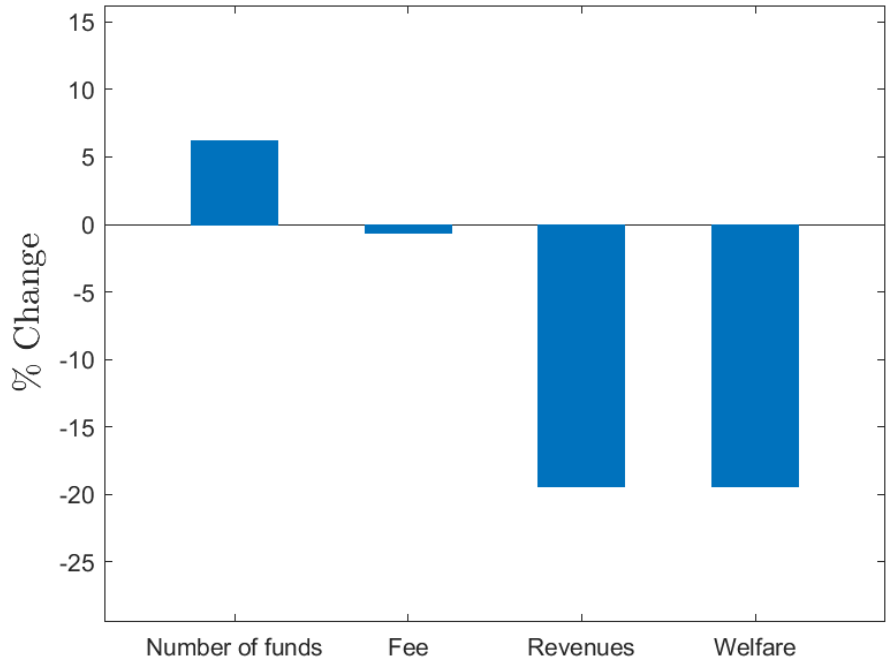

Figure 3.9: Percentage change in average number of funds, average fee, revenues and consumer surplus in the counterfactual where Blackrock is replaced by two firms identical to Charles Schwab. For number of funds, we report the percentage change in average number of funds, where the average is computed across time. For the fee, we report the percentage change in the average fee, where the average is computed across time. For revenues, we report the percentage change in average revenues, where the average is computed across time and management companies.

Replacing Blackrock with two less efficient firms would lead to a 6.23% increase in the average number of funds while leaving the average fee basically unaltered. While relatively more inefficient management companies can eventually substitute Blackrock funds, they can do so only over the long run while smoothing fund creation over time. This gradual substitution leads to a large a decrease in household surplus, equal to 19.46%. Thus, constraining efficient management companies in favor of competition is not necessarily the optimal solution for a policy maker interested in maximizing household welfare.

3.6.4 Asset pricing implications

In this section, we go back to the steady-state equilibrium that we are able to characterize analytically and, within this equilibrium, we use the estimates from section 3.6.2 to connect the competitive response of management companies to the evidence of inelastic demand for financial assets that the previous literature has documented. In particular, we estimate that a reduction in initiation costs c that induces a 1% increase in the steady-state wealth A increases the valuation of the equity index by 5.5%. In other words, the steady-state of our estimated model implies a multiplier ξ of household wealth on the equity index price equal to 5.5.

We start with the following proposition that provides a closed-form expression for the multiplier ξ in the steady-state of the model.

Proposition 8. *Under the conditions detailed in section 3.4.2, the steady-state multiplier ξ is given by*

$$\xi \equiv \frac{dP}{dA}\frac{A}{P} = \left(1 - \frac{1}{n(1+n)}\frac{1+\zeta(n)}{\zeta'(n)}\right) \tag{3.30}$$

with $\zeta(n) > 0$ and $\zeta'(n) < 0$ for $n > 1$.

Proof: See Appendix 3.8.

We use the estimates for $\{c_j\}_{j=0}^{5}$ derived and discussed in section 3.6.2 to compute $c = \sum_{j=0}^{5} c_j$. Moreover, we use equations (3.20), (3.21) and (3.22) to solve for the steady-state wealth A, index price P and number of funds n. Thus, we have all the inputs needed to produce an estimate for the steady-state multiplier ξ using equation (3.30). Details about the inputs used to estimate ξ are provided in Table 3.4. For completeness and to ease comparison, we reinclude in table 3.4 also parameters that have been already introduced but that enters the expression of ξ.

Parameter	Description	Value
c	Estimated initiation cost	0.18
P	Steady-state index price	0.54
A	Steady-state financial wealth	0.66
$\frac{D}{P}$	Steady-state dividend yield	0.03
n	Steady-state total number of funds	5.8
ξ	Steady-state multiplier	5.5

Table 3.4: Estimated multiplier

Our estimated steady-state multiplier of 5.5 was untargeted in the estimation, yet very aligned to estimates that the previous literature has reported. Among others, Gabaix and Koijen (2021) estimates a macro equity multiplier equal to 5 and shows that previous estimates range approximately between 1.5 and 5.5. Our estimate is thus consistent with previous work. Yet, to our knowledge, no previous work has microfounded the macro equity multiplier starting from the competitive dynamics of passive mutual funds and dominant management companies.

We conclude this section by performing a comparative static exercise in the steady-state of our model, where we vary the dollar dividend D and fund initiation costs c around the values reported in Table 3.4.

We start from varying c in Figure 3.10 and we flag in red the estimates we obtain in our model. The top left panel shows the steady state fee as function of c. Not surprisingly, the equilibrium fee increases with the initiation costs. From the perspective of our model, higher costs will push management companies to supply less funds. Lower competition in the mutual fund sector would then endogenously lead to higher fees. The top right panel looks at the equilibrium index price P and shows that as initiation costs rise, the equilibrium index price decreases. From the top left panel we know that higher initiation costs are passed-through investors via higher fees which in turn reduce household demand for the

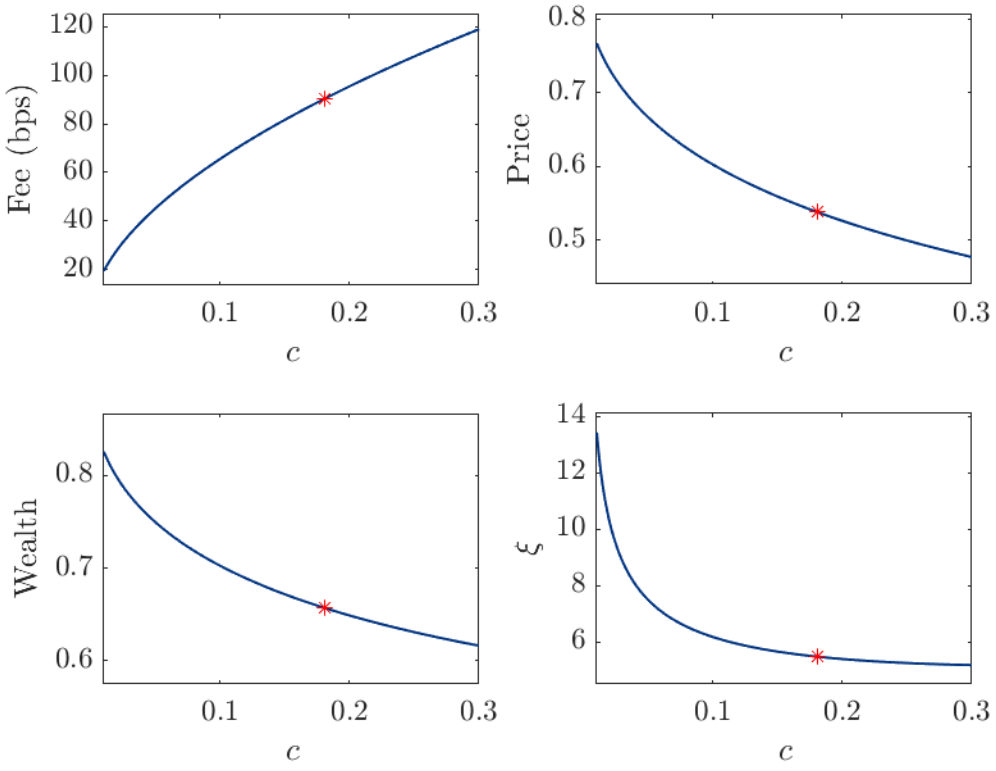

Figure 3.10: Equilibrium comparative static with respect to initiation costs c.

equity index. Finally, via market-clearing, lower demand for the equity index leads to a lower equilibrium price. Higher initiation costs also lead to lower equilibrium wealth A as shown in the bottom left panel. Once again, the mechanism for this outcome is driven by the competitive incentives in the mutual fund sector. Higher costs lead to lower fund creation and higher fees resulting in redistribution of wealth from household to mutual funds and management companies.

Lastly, the bottom right panel shows how the multiplier ξ varies with initiation costs. Increasing initiation costs from 0.01 to 0.3 decreases the steady-state multiplier from 13 to 4. To understand this result, consider first the case when c is small. As A increases, management companies find it optimal to create additional funds to collect part of the increase in wealth. Because c is small, management companies are able to initiate a large number of funds thereby significantly increasing competition in the mutual fund sector. As the equilibrium fee declines, household optimally invests a larger fraction of its wealth in the stock market, leading to an increase in the equity index price by market clearing. However, when c is higher, the creation of new funds becomes more and more costly for management companies. Thus, when A increases, management companies abstain from creating a large number of new funds. The result is that the level of competition in the mutual fund sector stays low, fees decline less and household increases less its demand for the equity index. The ultimate result is that, by market clearing, P increases but it increases less compared to the case of lower c.

Next, in Figure 3.11 we consider the comparative static of the same variables with respect to the dollar dividend D. As before, the top left panel shows the comparative static for the equilibrium fees. In this case, a higher dollar dividend leads to a decline in fees because a higher D increases the rate at which wealth accumulates. To accommodate the increase in asset demand, management companies create more funds. The stronger competition in the mutual fund sector ultimately leads to lower fees. Turning to the comparative static for P

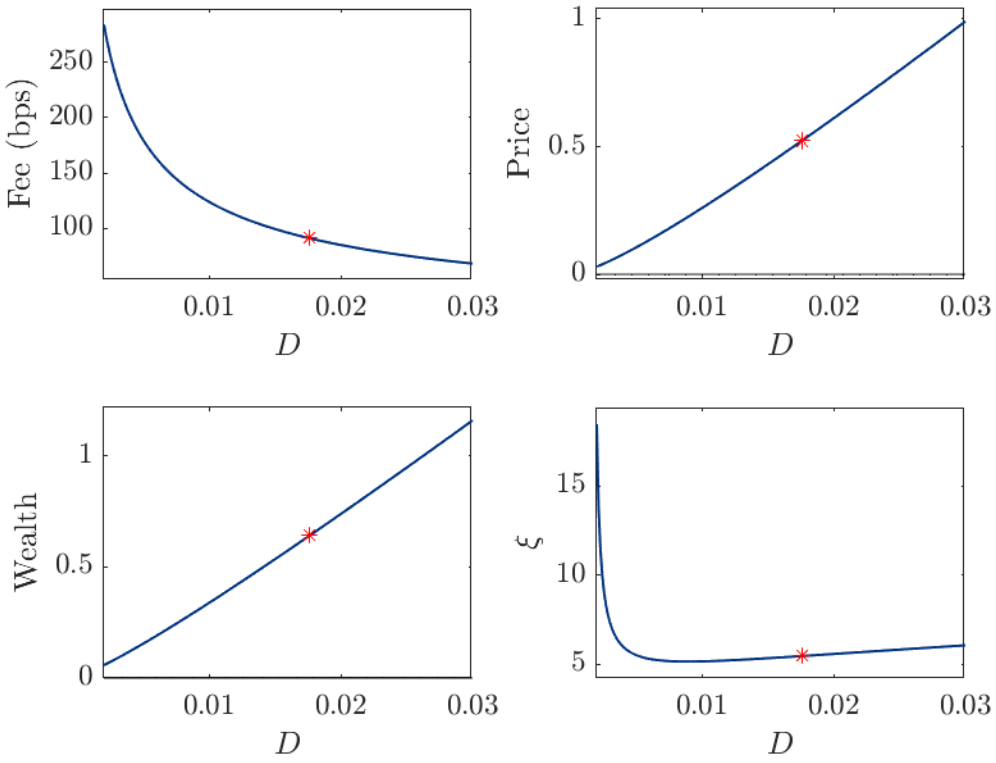

Figure 3.11: Equilibrium comparative static with respect to dividend D.

(top right panel) and A (bottom left), we notice that, differently from initiation costs, D affects the equilibrium price and wealth both directly and indirectly through the equilibrium number of funds n. Starting from the top right panel, we see that the index price increases with D. Indeed, a higher dividend increases the rate at which household wealth accumulates which in turn increases household demand for the equity index. Moreover, management companies accommodate the increase in demand by creating additional funds, leading to a decrease in fees and to a further increase in household demand. Both the direct effect on wealth as well as the indirect effect through n contribute to increasing household demand, ultimately leading to an increase in the index price through market clearing.

Turning to the bottom left panel, we can see that the equilibrium wealth increases with the dollar dividend D. The dividend affects the equilibrium wealth directly because it mechanically increases the rate at which wealth accumulates and indirectly through fund initiation. In other words, higher D directly increases wealth accumulation rate and indirectly prompts management companies to increase the number of funds, given the increase in demand. Stronger competition in the mutual fund sector leads to a decline in fees which further and indirectly accelerate wealth accumulation. This indirect effect is summarized by the term $\frac{1}{1+\zeta(n)}$ in equation (3.20). Because $\frac{1}{1+\zeta(n)}$ is an increasing function of n, it contributes to amplify the initial and direct increase in D.

Finally, the bottom right panel describes how the steady-state multiplier varies with D. Notice that the steady-state multiplier depends on D only indirectly, through n. Consider first the case of small D. In this case, household demand for the equity index is relatively low with the consequence that management companies are constrained to manage a relatively limited menu of funds. It follows however that any increase in household wealth is particularly attractive for management companies and they respond by creating to an increase in A by creating a higher number of funds compared to the case of high D. The larger response of management companies in turn leads to a larger decline in fee, a larger increase in household

demand and, ultimately, to a larger increase in the equity index price via market clearing.

3.7 Conclusions

In this paper, we propose a model where the competitive dynamics in the mutual fund industry are driven by the decisions of heterogeneous and multi-product management companies to initiate new funds. We provide sufficient conditions that guarantee existence and uniqueness of a steady state equilibrium characterized by a constant number of funds operated by each management company and a constant index price. In addition, we develop a numerical algorithm that solves for the equilibrium path of the number of funds created by each management company and the equilibrium market clearing asset price.

In the second part of the paper, we estimate the model using data on US passive equity funds that operate in the Large and Mid Cap sectors. For each of the five biggest management companies, we estimate their cost of initiating new funds and match the fund proliferation patterns observed in the data with the ones implied by our model. Moreover, to further validate the model, we show how the model implied time series of equilibrium fees closely follows the observed time series of average expense ratio.

With our estimated model parameters, we study several counterfactuals to understand the contribution of each management company to the secular decline in fees and the surplus of our household investors. In the first set of counterfactuals, we remove, one at the time, each of the top 5 management companies from the market. In all cases, investor surplus decreases substantially, although the magnitude is heterogeneous and depends on how efficient the removed company is. Removing the most efficient management company, Blackrock, reduces household welfare by 25%. In a second set of counterfactuals, we perform a similar exercise. However, instead of simply removing the most efficient company from the market, we replace it with two less efficient companies. Interestingly, investor surplus still goes down with a

reduction of about 20%. The key insight is that restricting efficient management companies to favor competition might ultimately hurt investor welfare.

Finally, we turn to the asset pricing implications of our model. Modelling competition across management companies in an asset market equilibrium framework allows us to microfound the price impact of large institutional investors through their technological primitives. A reduction in initiation costs pushes companies to create more funds, reducing equilibrium fees and increasing household asset demand for the equity index. The index price will then need to increase to clear this excess demand. Lastly, we derive a closed-form expression for the steady-state multiplier of the equity index price with respect to household wealth. Using our initiation cost estimates, we find that a 1% increase in household wealth implies a 5.5% increase in the steady-state index price. While this estimate is aligned with previous results in the literature, we microfound the equity multiplier through the competitive forces among large, heterogeneous and multiproduct investment companies.

3.8 Appendix: Derivations and Proofs

Derivation of HH portfolio allocation. Under the assumption of log utility, it is easy to verify that consuming a constant fraction of wealth is optimal for HH. In particular, from the Euler equation (3.4) and the budget constraint, one can verify that $C_t = (1-\beta)A_t$ is the optimal consumption in each period.

To derive the optimal portfolio allocation w_t, denote the log consumption and log wealth by $c_t \equiv \log(C_t)$ and $a_t \equiv \log(A_t)$ respectively so that $c_t = \log(1-\beta) + a_t$. The budget constraint in logs is then

$$\Delta a_{t+1} = \log(1 + w_t R_{t+1}) + \log(1-\beta) \tag{3.31}$$

$$\approx w_t r_{t+1} + \frac{1}{2} w_t (1-w_t)\sigma_t^2 + \log(1-\beta) \tag{3.32}$$

where $\Delta a_{t+1} \equiv a_{t+1} - a_t$, $r_{t+1} \equiv \log(1+R_{t+1})$ and the second line follows the log-linear approximation of log portfolio returns in Campbell and Viceira (2002). Next, note that under the assumption that r_{t+1} is a Gaussian stationary process, we can take logs on both sides of (3.4) to obtain

$$\mathbb{E}_t[\Delta c_{t+1}] = \log(\beta) + \rho_t - f_t + \frac{1}{2}\sigma_t^2 + \frac{1}{2}\mathbb{V}_t[\Delta c_{t+1}] - Cov_t[\Delta c_{t+1}, r_{t+1}].$$

Moreover, because we normalized the return on the risk-free to zero, the above expression boils down to

$$\mu_t - f_t + \frac{1}{2}\sigma_t^2 = Cov_t[\Delta c_{t+1}, r_{t+1}]. \tag{3.33}$$

Lastly, approximation (3.32) and the constant consumption-wealth ratio imply that we can

solve for w_t in (3.33) to obtain

$$w_t = \frac{\rho_t + \sigma_t^2/2 - f_t}{\sigma_t^2}. \tag{3.34}$$

Proof of Proposition 7. For given π_t, company j's Euler equation implied by problem (3.13) is given by

$$\begin{aligned} &\frac{\pi_t}{(1+n_t)^2} + \delta_j \beta \left(\frac{n_{jt+1} - n_{jt}}{n_{jt}} \right) \left[\frac{n_{jt+1}}{n_{jt}} + 1 \right] = \\ &\frac{2\pi_t n_{jt}}{(1+n_t)^3} + (1-\beta)c_j + \delta_j \left(\frac{n_{jt} - n_{jt-1}}{n_{jt-1}} \right) \end{aligned} \tag{3.35}$$

If a steady $\{(n_j)_{j=1}^M, P, A\}$ exists, then for given P and A, n_j must satisfy (3.35) which boils down to

$$n_j = \frac{1+n}{2} - \frac{(1-\beta)}{2\pi} c_j (1+n)^3 \tag{3.36}$$

where $\pi = \beta A \frac{\mu^2}{\sigma^2}$. Summing across j, the steady state total number of funds n in the market solves

$$\beta \frac{\mu^2}{\sigma^2} A(M + n(M-2)) = (1-\beta)(1+n)^3 c \tag{3.37}$$

Moreover, given the steady state fee

$$f = \frac{\mu}{n+1} \tag{3.38}$$

we can rewrite the equations that pin down the steady state P and A as

$$P = \frac{\mu n}{\sigma^2(1+n)}\beta A \tag{3.39}$$

$$A = \beta A + D - \frac{\mu}{1+n}P \tag{3.40}$$

where without loss of generality we normalized $\bar{Q} = 1$. From (3.39) and (3.40) we can solve for A and P as function of n and other parameters

$$P = \left(\frac{\frac{\mu}{\sigma^2}\frac{\beta}{1-\beta}\frac{n}{1+n}}{1+\zeta(n)} \right) D \tag{3.41}$$

$$A = \left(\frac{1}{1+\zeta(n)} \right) \frac{D}{1-\beta} \tag{3.42}$$

where

$$\zeta(n) \equiv \frac{\mu^2}{\sigma^2}\frac{\beta}{1-\beta}\frac{n}{(1+n)^2}. \tag{3.43}$$

The steady-state n can then be found by substituting (3.42) into (3.37)

$$\tilde{\pi}\left(\frac{1}{1+\zeta(n)} \right)(M + n(M-2)) = (1-\beta)(1+n)^3 c \tag{3.44}$$

which can be rearranged more conveniently as

$$\tilde{\pi}\,(M + n(M-2)) = (1-\beta)c\left[(1+n)^3 + \frac{\mu^2}{\sigma^2}\frac{\beta}{1-\beta}n(1+n) \right] \tag{3.45}$$

with $\tilde{\pi} \equiv \frac{D}{1-\beta}\frac{\beta\mu^2}{\sigma^2}$.

To show existence and uniqueness, note that at $n = 0$, the LHS of (3.45) is greater than its RHS provided $\tilde{\pi}M > (1-\beta)c$. Next, note that the LHS increases in n at a constant rate, whereas the RHS increases in n at an increasing rate. Thus, there will be one and only one

$n > 0$ at which (3.45) is satisfied ■

Proof of Proposition 8. Consider an increase in fund initiation costs c and note that the only way this change in costs affects the equilibrium wealth A and asset prices P is through the effect on n. Differentiating (3.41) and (3.39) with respect to n gives

$$\frac{dA}{dn} = -\frac{\zeta'(n)}{1+\zeta(n)}\frac{D}{1-\beta}$$
$$\frac{dP}{dn} = \frac{\mu}{\sigma^2}\frac{n}{1+n}\beta\frac{dA}{dn} + \frac{1}{(1+n)^2}\frac{\mu}{\sigma^2}\left(\frac{1}{1+\zeta(n)}\right)\frac{\beta}{1-\beta}D$$

Next, take the ratio of the two expressions above and note that

$$\frac{P}{A} = \frac{\mu}{\sigma^2}\beta\frac{n}{1+n} \tag{3.46}$$

we obtain

$$\frac{dP}{dA} = \frac{P}{A}\left(1 - \frac{1}{n(1+n)}\frac{1+\xi(n)}{\xi'(n)}\right) \tag{3.47}$$

with

$$\zeta'(n) = \frac{\mu^2}{\sigma^2}\frac{\beta}{1-\beta}\frac{1-n}{(1+n)^3} \tag{3.48}$$

which is negative for $n > 1$ ■

3.9 Appendix: Figures

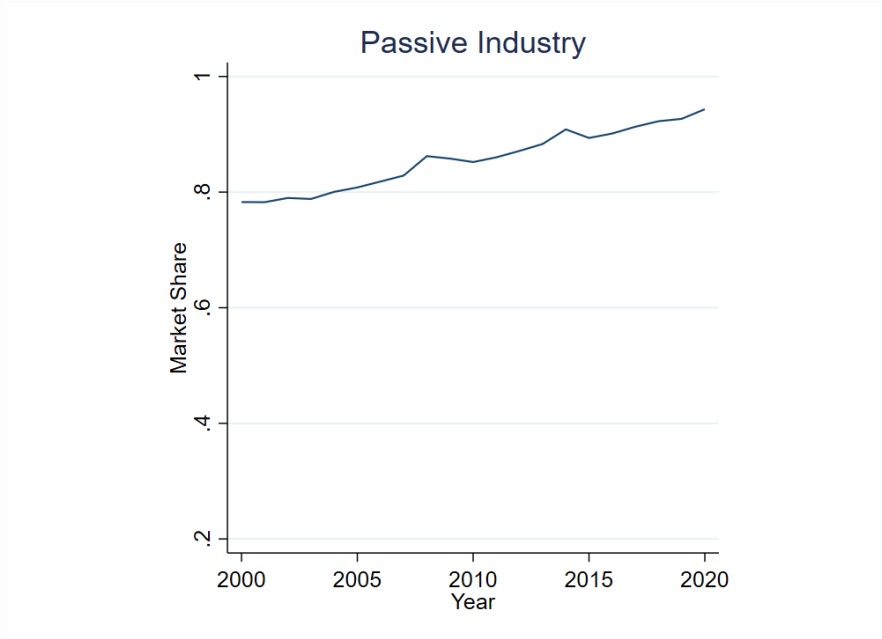

Figure 3.12: Market share of the five biggest investment companies in the passive industry. Market shares are in terms of end-of-year assets under management (AUM).

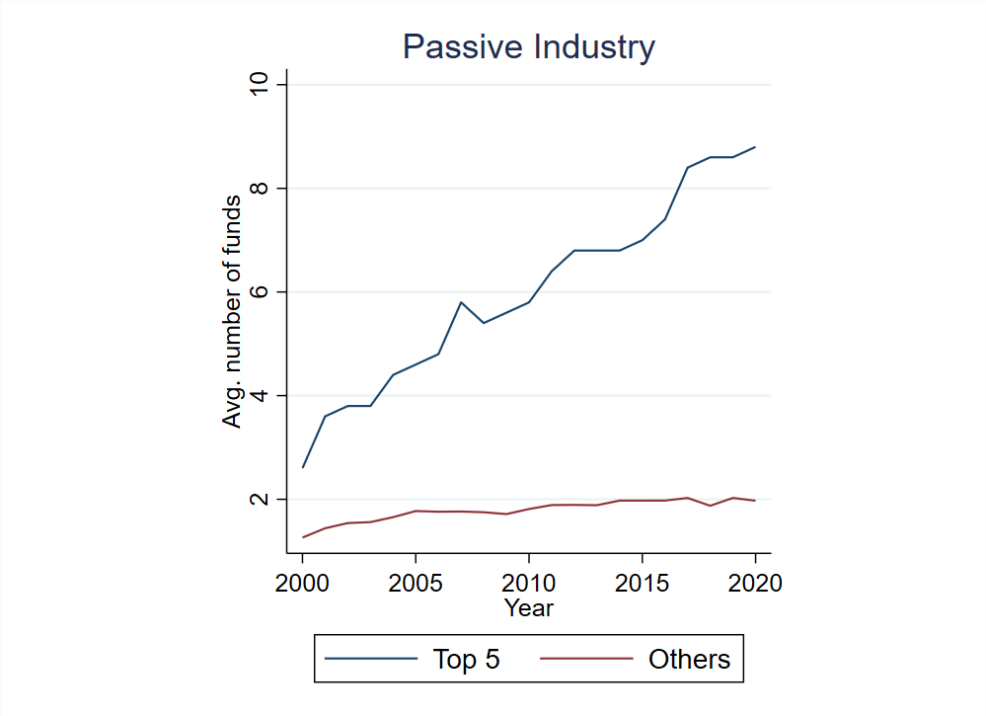

Figure 3.13: Average number of passive funds per management company. Funds with different share classes count as a single fund.

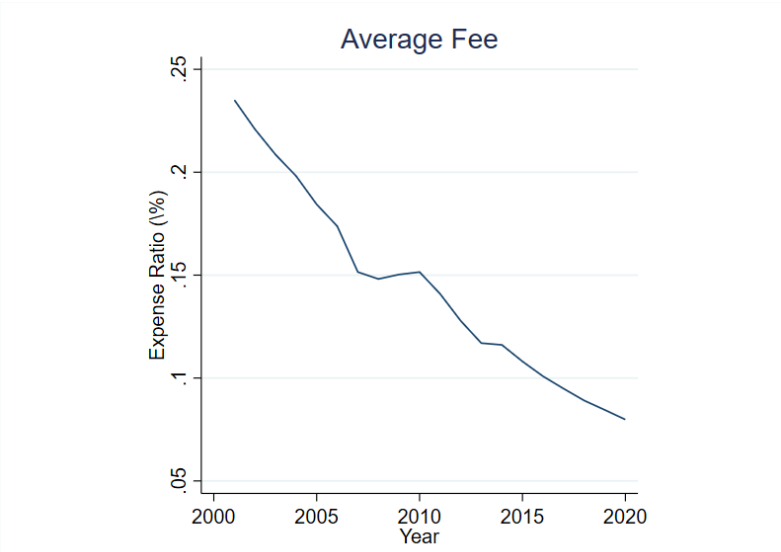

Figure 3.14: Average asset-weighted fee across passive funds. Funds with different share classes count as a single fund.

3.10 Appendix: Tables

	Obs.	Mean	Std. Dev	p5	p25	p50	p75	p95
AUM (bln.)	16552	2.00	13.41	0.02	0.08	0.28	0.96	5.76
Gross return (%)	16159	0.89	1.79	-2.52	-0.07	1.09	2.04	3.27
Expense Ratio (%)	16160	1.06	0.48	0.19	0.83	1.11	1.35	1.83
Passive	16552	0.22	0.42	0.00	0.00	0.00	0.00	1.00
Alpha (%)	13552	0.04	0.59	-0.47	-0.12	0.02	0.19	0.54
Market beta	13552	0.97	0.21	0.77	0.91	0.98	1.03	1.18
Market sharev(%)	16552	1.80	4.45	0.01	0.08	0.33	1.07	10.28
# of funds per company	16552	3.79	3.31	1.00	1.00	3.00	5.00	11.00

Table 3.5: Summary statistics of the full sample. All variables are winsorized at 1% and 99% levels. Returns and alpha are monthly. The expense ratio is annual.

	Obs.	Mean	Std. dev.	p5	p25	p50	p75	p95
AUM (bln.)	3697	5.70	27.72	0.02	0.09	0.42	1.89	18.10
Gross return (%)	3620	0.94	1.75	-2.12	-0.02	1.11	2.04	3.12
Expense Ratio	3621	0.48	0.42	0.08	0.20	0.35	0.60	1.57
Alpha (%)	3112	0.04	0.51	-0.31	-0.07	0.00	0.09	0.37
Market beta	3112	0.97	0.15	0.83	0.94	0.98	1.01	1.09
Market share (%)	3697	3.88	7.59	0.01	0.10	0.47	3.29	20.11
# of funds per company	3697	4.48	3.95	1.00	1.00	3.00	6.00	13.00

Table 3.6: Summary statistics for the passive sample. All variables are winsorized at 1% and 99% levels. Returns, alpha and expense ratios are monthly. The expense ratio is annual.

Company Code	Ticker	Fund Name	Beta	SMB	HML	MOM	Alpha	Gross Monthly Returns	Expense Ratio
VAN	VFIAX	Vanguard 500 Index Fund	0.9841	-0.1190	-0.0083	0.0002	-0.0001	-0.0028	0.0006
VAN	VINIX	Vanguard Institutional Index Fund	0.9864	-0.1196	-0.0082	0.0001	-0.0001	-0.0028	0.0005
SSB	SPY	SPDR S&P 500 ETF Trust	0.9863	-0.1222	-0.0057	0.0029	0.0000	-0.0028	0.0009
SSB	SSEYX	State Street Equity 500 Index II Portfolio	0.9830	-0.1251	-0.0141	0.0018	0.0003	-0.0028	0.0005
SSB	SVSPX	State Street S&P 500 Index Fund	0.9866	-0.1215	-0.0072	-0.0001	-0.0001	-0.0029	0.0016
SSB	SSSYX	State Street Equity 500 Index Fund	0.9889	-0.1157	-0.0066	0.0028	-0.0001	-0.0028	0.0013
BLK	IVV	iShares Core S&P 500 ETF	0.9864	-0.1200	-0.0081	0.0003	-0.0001	-0.0028	0.0005
BLK	WFSPX	iShares S&P 500 Index Fund	0.9858	-0.1191	-0.0079	-0.0003	-0.0001	-0.0028	0.0012
FID	FXAIX	Fidelity 500 Index Fund	0.9864	-0.1199	-0.0086	-0.0001	-0.0001	-0.0028	0.0005
CSW	SWPPX	Schwab S&P 500 Index Fund	0.9845	-0.1193	-0.0095	0.0003	-0.0001	-0.0028	0.0005
PRI	PREIX	T Rowe Price Equity Index 500 Fund	0.9855	-0.1194	-0.0083	-0.0002	-0.0001	-0.0028	0.0019
DFA	DFUSX	US Large Company Portfolio	0.9864	-0.1173	-0.0109	0.0000	-0.0001	-0.0028	0.0008
NTC	NOSIX	Stock Index Fund	0.9858	-0.1177	-0.0089	-0.0009	-0.0001	-0.0028	0.0010
USA	USPRX	S&P 500 Index Fund	0.9860	-0.1181	-0.0085	0.0000	-0.0001	-0.0028	0.0020
PGI	PLFIX	LargeCap S&P 500 Index Fund	0.9845	-0.1187	-0.0091	-0.0002	-0.0001	-0.0028	0.0027
DRY	DSPIX	Dreyfus Institutional S&P 500 Stock Index Fund	0.9857	-0.1206	-0.0063	0.0013	-0.0001	-0.0028	0.0028
DRY	PEOPX	Dreyfus S&P 500 Index Fund	0.9873	-0.1200	-0.0075	0.0009	-0.0001	-0.0028	0.0050
TIA	TISPX	S&P 500 Index Fund	0.9854	-0.1202	-0.0072	-0.0004	-0.0001	-0.0028	0.0011
SEI	SPINX	S&P 500 Index Fund	0.9853	-0.1167	-0.0055	0.0023	-0.0001	-0.0028	0.0005
SEI	SSPIX	S&P 500 Index Fund	0.9856	-0.1192	-0.0085	0.0003	-0.0001	-0.0028	0.0028
JPM	OGFAX	JPMorgan Equity Index Fund	0.9862	-0.1202	-0.0085	0.0005	-0.0001	-0.0028	0.0017
LBR	NINDX	Columbia Large Cap Index Fund	0.9887	-0.1206	-0.0081	0.0006	-0.0001	-0.0029	0.0026
GWC	MXVIX	Great-West S&P 500 Index Fund	1.0046	0.3668	0.2179	0.0117	0.0031	-0.0033	0.0040
MAS	MMIZX	MM S&P 500 Index Fund	0.9880	-0.1200	-0.0083	0.0001	-0.0001	-0.0028	0.0039
NFS	GRMIX	Nationwide S&P 500 Index Fund	0.9863	-0.1219	-0.0089	-0.0013	-0.0001	-0.0028	0.0030
DWS	SCPIX	DWS S&P 500 Index Fund	0.9792	-0.1191	-0.0122	-0.0001	0.0000	-0.0025	0.0045
DWS	BTIEX	DWS Equity 500 Index Fund	0.9790	-0.1204	-0.0109	0.0000	0.0000	-0.0025	0.0027
WFB	WFILX	Wells Fargo Index Fund	0.9864	-0.1196	-0.0085	0.0003	-0.0001	-0.0028	0.0038
AIM	SPIAX	Invesco S&P 500 Index Fund	0.9851	-0.1187	-0.0071	0.0008	-0.0001	-0.0028	0.0072
ABF	GEQYX	Equity Index Fund	0.9904	-0.1039	-0.0027	-0.0020	0.0000	-0.0026	0.0012

Table 3.7: Top 30 passive funds in the Large Cap sector.

Company Code	Ticker	Fund Name	Beta	SMB	HML	MOM	Alpha	Gross Monthly Returns	Expense Ratio
VAN	VIMAX	Vanguard Mid-Cap Index Fund	0.9266	0.1544	-0.1520	-0.0888	-0.0016	-0.0070	0.0005
VAN	VEXAX	Vanguard Extended Market Index Fund	0.9793	0.5387	-0.0685	0.0017	-0.0010	-0.0068	0.0006
VAN	VOE	Vanguard Mid-Cap Value Index Fund	0.9127	0.0971	0.0370	-0.1222	-0.0015	-0.0100	0.0007
VAN	VMGMX	Vanguard Mid-Cap Growth Index Fund	0.9390	0.2126	-0.3615	-0.0567	-0.0017	-0.0036	0.0007
VAN	VSPMX	Vanguard S&P Mid-Cap 400 Index Fund	0.9550	0.4006	0.0800	0.0367	-0.0006	-0.0085	0.0010
VAN	IVOG	Vanguard S&P Mid-Cap 400 Growth Index Fund	0.9667	0.3504	-0.0314	0.1663	-0.0008	-0.0077	0.0015
VAN	IVOV	Vanguard S&P Mid-Cap 400 Value Index Fund	0.9408	0.4493	0.1865	-0.1036	-0.0005	-0.0093	0.0015
BLK	IJH	iShares Core S&P Mid-Cap ETF	0.9623	0.3957	0.0776	0.0412	-0.0005	-0.0084	0.0008
BLK	IWR	iShares Russell Mid-Cap ETF	0.9243	0.1976	-0.0896	-0.0579	-0.0014	-0.0068	0.0019
BLK	IWS	iShares Russell Mid-Cap Value ETF	0.8752	0.1899	0.0787	-0.0871	-0.0016	-0.0099	0.0024
BLK	IWP	iShares Russell Mid-Cap Growth ETF	0.9805	0.2031	-0.2968	-0.0121	-0.0009	-0.0028	0.0024
BLK	IJK	iShares S&P Mid-Cap 400 Growth ETF	0.9685	0.3477	-0.0329	0.1651	-0.0010	-0.0079	0.0024
BLK	IJJ	iShares S&P Mid-Cap 400 Value ETF	0.9413	0.4490	0.1859	-0.1039	-0.0005	-0.0093	0.0025
BLK	BRMKX	iShares Russell Mid-Cap Index Fund	0.9267	0.1785	-0.0932	-0.1354	-0.0017	-0.0068	0.0015
BLK	JKG	iShares Morningstar Mid-Cap ETF	0.9659	0.1216	-0.0766	-0.0828	-0.0026	-0.0099	0.0025
BLK	JKI	iShares Morningstar Mid-Cap Value ETF	0.8690	0.1603	0.2226	-0.1096	-0.0012	-0.0101	0.0030
BLK	JKH	iShares Morningstar Mid-Cap Growth ETF	0.9811	0.2746	-0.3795	-0.0104	-0.0012	-0.0015	0.0030
BLK	BSMKX	iShares Russell Small/Mid-Cap Index Fund	1.0170	0.5240	0.0045	-0.0239	-0.0015	-0.0075	0.0013
BLK	SMMD	iShares Russell 2500 ETF	1.1511	0.0000	0.0000	0.0000	-0.0018	-0.0452	0.0006
FID	FSMAX	Fidelity Extended Market Index Fund	0.9792	0.5365	-0.0687	0.0034	-0.0010	-0.0069	0.0005
FID	FSMDX	Fidelity Mid Cap Index Fund	0.9233	0.1932	-0.0919	-0.0591	-0.0013	-0.0069	0.0005
FID	FZFLX	Fidelity SAI Small-Mid Cap 500 Index Fund	0.9554	0.3259	-0.0766	-0.0718	-0.0017	-0.0067	0.0014
SSB	MDY	SPDR S&P MidCap 400 ETF	0.9527	0.3999	0.0794	0.0363	-0.0006	-0.0085	0.0024
SSB	MDYG	SPDR S&P 400 Mid Cap Growth ETF	1.0514	0.4881	0.1740	0.0753	0.0000	-0.0077	0.0015
SSB	SPMD	SPDR Portfolio S&P 400 Mid Cap ETF	1.0051	0.5495	0.0771	0.0314	-0.0012	-0.0076	0.0006
SSB	MDYV	SPDR S&P 400 Mid Cap Value ETF	0.9637	0.6339	0.2154	-0.2250	0.0013	-0.0093	0.0015
SSB	SSMHX	State Street Small/Mid Cap Equity Index Portfolio	1.0035	0.5111	-0.0461	-0.0204	-0.0011	-0.0066	0.0005
SSB	SSMKX	State Street Small/Mid Cap Equity Index Fund	0.9645	0.5212	-0.0466	0.0150	-0.0012	-0.0066	0.0008
DFA	DFFVX	US Targeted Value Portfolio	1.0049	0.7027	0.3546	-0.0201	-0.0013	-0.0125	0.0037
CSW	SCHM	Schwab US Mid-Cap ETF	0.9576	0.3030	-0.0782	-0.0053	-0.0005	-0.0064	0.0005

Table 3.8: Top 30 passive funds in the Mid Cap sector.

REFERENCES

Abis, S. and Lines, A. (2022), 'Do mutual funds keep their promises?', SSRN working paper.

Agarwal, S., Grigsby, J., Hortaçsu, A., Matvos, G., Seru, A. and Yao, V. (2021), 'Searching for approval', working paper.

Aguirregabiria, V., Collard-Wexler, A. and Ryan, S. (2021), 'Dynamic games in empirical industrial organization', NBER working paper.

Allen, J., Clark, R. and Houde, J.-F. (2014), 'The effect of mergers in search markets: Evidence from the canadian mortgage industry', *American Economic Review* 104(10), 3365–3396.

Amir, R. and Jin, J. (2001), 'Cournot and bertrand equilibria compared: substitutability, complementarity and concavity', *International Journal of Industrial Organization* 19, 303–317.

Andrews, I., Gentzkow, M. and Shapiro, J. (2017), 'Measuring the sensitivity of parameter estimates to estimation moments', *Quarterly Journal of Economics* 132(4), 1533–1592.

Anton, M., Ederer, F., Gine, M. and Schmalz, M. (2017), 'Common ownership, competition, and top management incentives', *Journal of Political Economy* 81(3), 669–728.

Azar, J., Schmalz, M. and Tecu, I. (2018), 'Anticompetitive effects of common ownership', *Journal of Finance* 73(4), 1513–1565.

Azar, J. and Vives, X. (2021), 'General equilibrium oligopoly and ownership structure', *Econometrica* 89(3), 999–1048.

Backus, M., Conlon, C. and Sinkinson, M. (2021), 'Common ownership and competition in the ready-to-eat cereal industry', NBER working paper.

Badarinza, C., Campbell, J. Y. and Ramadorai, T. (2016), 'International comparative household finance', *Annual Review of Economics* 8(1), 111–144.

Badoer, D., Costello, C. and James, C. (2020), 'I can see clearly now: The impact of disclosure requirements on 401(k) fees', *Journal of Financial Economics* 136, 471–489.

Bai, J., Philippon, T. and Savov, A. (2016), 'Have financial markets become more informative', *Journal of Financial Economics* 122, 625–654.

Baker, M., Egan, M. and Sarkar, S. (2022), 'How do investors value esg?', NBER working paper.

Balduzzi, P. and Reuter, J. (2018), 'Heterogeneity in target date funds: Strategic risk-taking or risk matching?', *The Review of Financial Studies* 32(1), 300337.

Ballester, C., Calvó-Armenagol, A. and Zenou, Y. (2006), 'Who's who in networks. wanted: The key player', *Econometrica* 74(5), 1403–1417.

Barberis, N. and Shleifer, A. (2003), 'Style investing', *Journal of Financial Economics* 68(2), 161–199.

Baruch, S. and Zhang, X. (2022), 'The distortion in prices due to passive investing', *Management Science* forthcoming.

Basak, S. and Pavlova, A. (2013), 'Asset prices and institutional investors', *American Economic Review* 103(5), 1728–1758.

Beggs, S., Cardell, S. and Hausman, J. (1981), 'Assessing the potential demand for electric cars', *Journal of Econometrics* 16, 1–19.

Ben-David, I., Franzoni, F., Kim, B. and Moussawi, R. (2022), 'Competition for attention in the etf space', *Review of Finacial Studies* forthcoming.

Benartzi, s. and Thaler, R. (2001), 'Naive diversification strategies in defined contribution saving plans.', *American Economic Review* 91(1), 2997–3053.

Benartzi, s. and Thaler, R. (2007), 'Heuristics and biases in retirement savings behavior.', *Journal of Economic Perspectives* 21(3), 81–104.

Benetton, M. (2021), 'Leverage regulation and market structure: A structural model of the u.k. mortgage market', *The Journal of Finance* 76(6), 2997–3053.

Benetton, M., Buchak, G. and Robles-Garcia, C. (2022), 'Wide or narrow? competition and scope in financial intermediation', working paper.

Berk, J., Binsbergen, J. and Liu, B. (2017), 'Matching capital and labor', *Journal of Finance* 72(6), 500–523.

Berk, J. and Green, R. (2004), 'Mutual fund flows and performance in rational markets', *Journal of Political Economy* 112(6), 1269–1295.

Berry, S. (1994), 'Estimating discrete-choice models of product differentiation', *The RAND Journal of Economics* 25(2), 242–262.

Berry, S., Levinshon, J. and Pakes, A. (1999), 'Voluntary export restraints on automobiles: Evaluating a trade policy', *American Economic Review* 89(3), 400–430.

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

Beshears, J., Choi, J., Laibson, D. and Madrian, B. (2009), 'The importance of default options for retirement saving outcomes: Evidence from the united states', *The Importance of Default Options for Retirement Saving Outcomes: Evidence from the United States.* University of Chicago Press.

Betermier, S., Schumacher, D. and Shahrad, A. (2022), 'Menu proliferation and entry deterrence', *Review of Asset Pricing Studies* forthcoming.

Bhattacharya, U., Lee, J. and Pool, V. (2013), 'Conflicting family values in mutual fund families', *Journal of Finance* 68(1), 173–200.

Bhattacharya, V., Illanes, G. and Padi, M. (2020), 'Fiduciary duty and the market for financial advice', working paper.

Bhattacharya, V. and Illannes, G. (2022), 'The design of defined contributions plans', NBER working paper.

Bonacich, P. (1987), 'Power and centrality: A family of measures', *American Journal of Sociology* 92(5), 1170–1182.

Bond, P. and Garcia, D. (2022), 'The equilibrium consequences of indexing', *Review of Financial Studies* 35(7), 31753230.

Brancaccio, G., Li, D. and Schüroff, N. (2020), 'Learning by trading: The case of the us market for municipal bonds', working paper.

Bresnahan, T. (1987), 'Competition and collusion in the american automobile industry: The 1955 price war.', *The Journal of Industrial Organization* 35(4), 457–481.

Brown, D. and Davies, S. (2021), 'Off target: On the underperformance of target-date funds', SSRN working paper.

Buchack, G., Matvos, G., Piskorski, T. and Seru, A. (2022), 'Beyond the balance sheet model of banking: Implications for bank regulation and monetary policy', NBER working paper.

Buchak, G., Matvos, G., Piskorski, T. and Seru, A. (2018), 'Fintech, regulatory arbitrage, and the rise of shadow banks', *Journal of Financial Economics* 130(3), 453–483.

Campbell, J. and Viceira, L. (2002), 'Strategic asset allocation: Portfolio choice for long-term investors', Oxford University Press.

Carhart, M. M. (1997), 'On persistence in mutual fund performance', *The Journal of Finance* 52(1), 57–82.

Carroll, G., Choi, J., Laibson, D., Madrian, B. and Metrick, A. (2009), 'Optimal defaults and active decisions', *The Quarterly Journal of Economics* 124(4), 16391674.

Chalmers, J. and Reuter, J. (2020), 'Is conflicted investment advice better than no advice?', *Journal of Financial Economics* 138(2), 366387.

Chen, Y., Zenou, Y. and Zhou, J. (2018), 'Competitive pricing strategies in social networks', *The RAND Journal of Economics* 49(3), 672705.

Chen, Y., Zenou, Y. and Zhou, J. (2022), 'The impact of network topology and market structure on pricing', *Journal of Economic Theory* 204.

Cheng, L. (1985), 'Comparing bertrand and cournot equilibria: A geometric approach', *RAND Journal of Economics* 16(1), 146–152.

Chevalier, J. and Ellison, G. (1997), 'Risk taking by mutual funds as a response to incentives', *Journal of Political Economy* 105(6), 1167–1200.

Choi, J. (2015), 'Contributions to defined contribution pension plans', *Annual Reviews of Financial Economics* 7, 161–178.

Choukhmane, T. (2021), 'Default options and retirement saving dynamics', working paper.

Coles, J., Heath, D. and Ringgenberg, M. (2022), 'On index investing', *Journal of Financial Economics* forthcoming.

Conlon, C. and Gortmaker, J. (2020), 'Best practices for differentiated products demand estimation with pyblp', *RAND Journal of Economics* 51(4), 1108–1161.

Cuesta, I. and Sepúlveda, A. (2020), 'Price regulation in credit markets: A trade-off between consumer protection and credit access', working paper.

Dermine, J., Neven, D. and Thisse, J. (1991), 'Towards an equilibrium model of the mutual funds industry', *Journal of Banking and Finance* 15, 485–499.

Duarte, V., Fonseca, J., Goodman, A. and Parker, J. (2022), 'Simple allocation rules and optimal portfolio choice over the lifecycle', NBER working paper.

Dubois, P., Griffith, R. and Nevo, A. (2014), 'Do prices and attributes explain international differences in food purchases?', *American Economic Review* 104(3), 832–867.

Ederer, F. and Pellegrino, B. (2022), 'A tale of two networks: Common ownership and product market rivalry', SSRN working paper.

Egan, M. (2019), 'Brokers versus retail investors: Conflicting interests and dominated products', *The Journal of Finance* 74(3), 1217–1260.

Egan, M., Hortaçsu, A. and Matvos, G. (2017), 'Deposit competition and financial fragility: Evidence from the us banking sector', *American Economic Review* 107(1), 169–216.

Egan, M., MacKay, A. and Yang, H. (2023), 'What drives variation in investor portfolios? estimating the roles of beliefs and risk preferences.', NBER working paper.

Epple, D. (1987), 'Hedonic prices and implicit markets: Estimating demand and supply functions for differentiated products', *Journal of Political Economy* 95(1), 5980.

Ericson, R. and Pakes, A. (1995), 'Markov-perfect industry dynamics: A framework for empirical work', *Review of Economics Studies* 62(1), 53–82.

Fama, E. and French, K. (1992), 'The cross-section of expected stock returns', *The Journal of Finance* 47(2), 427–465.

Fama, E. and Kenneth, F. (2010), 'Luck versus skill in the cross-section of mutual fund returns', *The Journal of Finance* 65(5), 1915–1947.

Farboodi, M., Matray, A., Veldkamp, L. and Venkateswaran, V. (2021), 'Where has all the data gone?', *Review of Financial Studies* 35(7), 31013138.

Feenstra, R. and Levinsohn, J. (1995), 'Estimating markups and market conduct with multidimensional product attributes', *Review of Economic Studies* 62(1), 19–52.

Gabaix, X. and Koijen, R. (2021), 'In search of the origins of financial fluctuations: the inelastic markets hypothesis', NBER working paper.

Galeotti, A., Golub, B., Goyal, S., Talamàs, E. and Tamuz, O. (2022), 'Taxes and market power: A principal components approach', arXiv prepeint arXiv:2112.08153v2.

Gandhi, A. and Houde, J. (2023), 'Measuring substitution patterns in differentiated-products industries', NBER working paper.

Gaspar, J., Massa, M. and Matos, P. (2006), 'Favoritism in mutual fund families? evidence on strategic cross-fund subsidization', *Journal of Finance* 61(1), 73–104.

Gentzkow, M. (2007), 'Valuing new goods in a model with complementarity: Online newspapers', *American Economic Review* 97(3), 713744.

Gil-Bazo, J. and Ruiz-Verdu, P. (2009), 'The relation between price and performance in the mutual fund industry', *The Journal of Finance* 64(5), 2153–2183.

Goeree, M. (2008), 'Limited information and advertising in the u.s. personal computer industry', *Econometrica* 16, 1017–1074.

Grice, R. (2023), 'Chain-linked markets', working paper.

Grice, R. and Gueciouer, A. (2023), 'Mutual fund market structure and company fee competition: Theory and evidence', SSRN working paper.

Gropper, M. (2023), 'Lawyers setting the menu: The effects of litigation risk on employer-sponsored retirement plans', working paper.

Gruber, M. (1996), 'Another puzzle: The growth in actively managed mutual funds', *The Journal of Finance* 51(3), 783–810.

Haddad, V., Huebner, P. and Loualiche, E. (2022), 'How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing'.

Hausman, J. (1996), 'Valuation of new goods under perfect and imperfect competition.', *The Economics of New Goods* pp. 207–248.

He, Z. and Krishnamurthy, A. (2013), 'Intermediary asset pricing', *American Economic Review* 103(2), 732–770.

Hoberg, G. and Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy* 124(5), 1423–1465.

Hortaçsu, A. and McAdams, D. (2010), 'Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the turkish treasury auction market', *Journal of Political Economy* 118(5), 833–865.

Hortaçsu, A. and Syverson, C. (2004), 'Product differentiation, search costs, and competition in the mutual fund industry: A case study of s&p 500 index funds.', *Quarterly Journal of Economics* 119(2), 403–456.

Hotelling, H. (1929), 'Stability in competition', *The Economic Journal* 39(153), 41–57.

Huberman, G. and Jiang, W. (2006), 'Offering versus choice in 401(k) plans: Equity exposure and number of funds', *Journal of Finance* 61(2), 763–801.

Inderst, R. and Ottaviani, M. (2012a), 'Competition through commissions and kickbacks', *American Economic Review* 102(2), 780809.

Inderst, R. and Ottaviani, M. (2012b), 'How (not) to pay for advice?', *Journal of Financial Economics* 105, 393–411.

Jackson, M. (2008), *Social and Economic Networks*, Princeton University Press.

Jensen, M. C. (1968), 'The performance of mutual funds in the period 1945–1964', *The Journal of Finance* 23(2), 389–416.

Kacperczyk, M., Nosal, J. and Sundaresan, S. (2022), 'Market power and price informativeness', *Review of Economic Studies* forthcoming.

Kacperczyk, Sialm, C. and Zheng, L. (2005), 'On the industry concentration of actively managed equity mutual funds', *The Journal of Finance* 60(4), 1983–2011.

Kastl, J. (2011), 'Discrete bids and empirical inference in divisible good auctions', *The Review of Economic Studies* 78(3), 974–1014.

Khorana, A. and Servaes, H. (1999), 'The determinants of mutual fund starts', *The Review of Financial Studies* 12(5), 1043–1074.

Koijen, R. and Yogo, M. (2019), 'A demand system approach to asset pricing', *Journal of Political Economy* 127(4), 1475–1515.

Koijen, R. and Yogo, M. (2022), 'The fragility of market risk insurance', *The Journal of Finance* 77(2), 815–862.

Kostovetsky, L. and Warner, J. (2020), 'Measuring innovation and product differentiation: Evidence from mutual funds', *Journal of Finance* 75(2), 779–823.

Kronlund, M., Pool, V., Sialm, C. and Stefanescu, I. (2021), 'Out of sight no more? the effect of fee disclosures on 401(k) investment allocations', *Journal of Financial Economics* 141, 644–668.

Lee, S. and Allenby, G. (2009), 'A direct utility model for market basket data', working paper.

Loseto, M. (2023), 'Network games of imperfect competition: An empirical framework', SSRN working paper.

Ma, L., Tang, Y. and Gomez, J. (2019), 'Portfolio manager compensation in the u.s. mutual fund industry', *Journal of Finance* 74(2), 587–638.

Madrian, B. and Shea, D. (2001), 'The power of suggestion: Inertia in 401(k) participation and savings behavior', *The Quarterly Journal of Economics* 116(4), 1149–1187.

Magnolfi, L., McClure, J. and Sorensen, A. (2022), 'Triplet embeddings for demand estimation', SSRN working paper.

Magnolfi, L., Quint, D., Sullivan, C. and Waldfogel, S. (2022), 'Differentiated-products cournot attributes higher markups than bertrandnash', *Economics Letters* 219, 166–175.

Malikov, G. (2021), 'Information, partecipation and passive investing', working paper.

Markowitz, H. (1952), 'Portfolio selection', *Journal of Finance* 7(1), 77–91.

Maskin, E. and Tirole, J. (1988), 'A theory of dynamic oligopoly, i: Overview and quantity competition with large fixed costs', *Econometrica* 56(3), 549–599.

Massa, M. (2003), 'How do family strategies affect fund performance? when performance-maximization is not the only game in town', *Journal of Financial Economics* 67, 249–304.

Mellman, G. and Sanzenbacher, G. (2018), '401(k) lawsuits: What are the causes and consequences?', Center for Retirement Research, Boston College(18-8).

Nanda, V., Narayanan, M. and Warther, V. (2000), 'Liquidity, investment ability, and mutual fund structure', *Journal of Financial Economics* 57, 417–443.

Nelson, S. (2020), 'Private information and price regulation in the us credit card market', working paper.

Nevo, A. (2001), 'Measuring market power in the ready-to-eat cereal industry', *Econometrica* 69(2), 307342.

Okuguchi, K. (1987), 'Equilibrium prices in the bertrand and cournot oligopolies', *Journal of Economic Theory* 42, 128–139.

Pástor, P., Stambaugh, R. and Taylor, L. (2017), 'Do funds make more when they trade more?', *Journal of Finance* 72(4), 1483–1528.

Pástor, P., Stambaugh, R. and Taylor, L. (2020), 'Fund tradeoffs', *Journal of Financial Economics* 138, 614–634.

Pavlova, A. and Sikorskaya, T. (2022), 'Benchmarking intensity', *Review of Financial Studies* forthcoming.

Pellegrino, B. (2023), 'Product differentiation and oligopoly: A network approach', SSRN working paper.

Petajisto, A. (2009), 'Why do demand curves for stocks slope down?', *Journal of Financial and Quantitative Analysis* 44(5), 1013–1044.

Pool, V., Sialm, C. and Stefanescu, I. (2016), 'It pays to set the menu: Mutual fund investment options in 401(k) plans', *The Journal of Finance* 521(4), 1179–1812.

Pool, V., Sialm, C. and Stefanescu, I. (2022), 'Mutual fund revenue sharing in 401k plans', NBER working paper.

Posner, E., Scott Morton, F. and Weyl, G. (2017), 'A proposal to limit the anticompetitive power of institutional investors', *Antitrust Law Journal* 81(3), 669–728.

Reuter, J. and Richardson, D. (2022), 'New evidence on the demand for advice within retirement plans', working paper.

Richert, E. (2021), 'Quantity commitment in multiunit auctions: Evidence from credit event auctions', working paper.

Robles-Garcia, C. (2021), 'Competition and incentives in mortgage markets: The role of brokers', working paper.

Rosen, S. (1974), 'Hedonic prices and implicit markets: Product differentiation in pure competition', *Journal of Political Economy* 82(1), 3455.

Roussanov, N., Ruan, H. and Wei, Y. (2021), 'Marketing mutual funds', *Review of Financial Studies* 34, 3045–3094.

Sammon, M. (2022), 'Passive ownership and price informativeness', SSRN working paper.

Sandhya, V. (2011), 'Agency problems in target date funds.', working paper.

Schmalz, M. and Zame, W. (2023), 'Index funds, asset prices, and the welfare of investors', SSRN working paper.

Singh, N. and Vives, X. (1984), 'Price and quantity competition in a differentiated duopoly', *RAND Journal of Economics* 15(4), 546–554.

Sirri, E. and Tufano, P. (1998), 'Costly search and mutual fund flows', *Journal of Finance* 53(5), 1589–1622.

Stigler, G. (1961), 'The economics of information', *Journal of Political Economy* 69(3), 213–225.

Tang, N., Mitchell, O., Mottola, G. and Utkus, S. (2010), 'The efficiency of sponsor and participant portfolio choices in 401(k) plans', *Journal of Public Economics* 94(2), 1073–1085.

Thomassen, ., Smith, H., Seiler, S. and Schiraldi, P. (2017), 'Multi-category competition and market power: A model of supermarket pricing', *American Economic Review* 107(8), 23082351.

Ushchev, P. and Zenou, Y. (2018), 'Price competition in product variety networks', *Games and Economic Behavior* 110(2), 226–247.

Vanguard (2022), 'How america saves.'.

Vives, X. (1985), 'On the efficiency of bertrand and cournot equilibria with product differentiation', *Journal of Economic Theory* 36, 166–175.

Weintraub, G., Benkard, L. and Van Roy, B. (2008), 'Markov perfect industry dynamics with many firms', *Econometrica* 76(6), 1375–1411.

Wermers, R. (2000), 'Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses', *The Journal of Finance* 55(4), 1655–1695.

Yang, H. (2023), 'What determines 401(k) plan fees? a dynamic model of transaction costs and markups', working paper.

Ørpetveit, A. (2021), 'Competition and fund family product development', SSRN working paper.