Essays on Social Demand Estimation: Evidence from CAPS and Steam

Andrew Clayton

A dissertation submitted in partial fulfillment

of the requirements for the degree of:

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Elaina Rose, Chair

Pat Bajari

Dong-Jae Eun

Stephan Siegel

Program Authorized to Offer Degree:

Department of Economics

University of Washington

Abstract

Essays on Social Demand Estimation:
Evidence from CAPS and Steam

Andrew Clayton

Chair of the Supervisory Committee:

Associate Professor Elaina Rose

Department of Economics

This dissertation examines the role of peer influence on the demand decisions made by socially connected consumers within markets for household investment and digital video games. In both applications, novel and rich data provides visibility into demand outcomes and the evolution of social networks, allowing for the estimation of demand models which quantify the impact of socialization on individual demand. These findings contribute to our understanding of how increases in connectivity and visibility into the consumption choices of peers change market function and consumer choice.

Chapter 1: Evidence suggests that the investment choices of household investors are influenced by the financial decisions made by their peers. An unresolved empirical challenge is to quantify the impact of socialization on portfolio composition relative to other factors. Rich individual data from the Motley Fool CAPS platform provides an opportunity to link asset selection decisions through an observable social network of peers. I collect data for nearly one-thousand interconnected users who collectively produce over 600,000 stock predictions between 2007 and 2013 and estimate the likelihood of asset demand conditional on a rich set of financial, macroeconomic, and individual factors. I further augment this model with measures of network socialization and peer influence to characterize how the decisions made by a user's peer group affect their own investment outcomes. The results demonstrate that the

inclusion of social covariates in a model for asset demand substantially improves explanatory power and that social factors appear strongly influential in explaining household investor actions in contrast to traditional financial factors.

Chapter 2: The PC video gaming industry, currently valued at over $30 billion, features customers whose demand decisions are motivated by a variety of factors including price, perceived quality, and software attributes. Gamers are highly social and sensitive to the opinions of reviewers, peers in the gaming community, and most notably their own immediate social groups. Using novel and richly descriptive customer-level data from Steam, a leading player in the digital distribution space, this paper presents demand estimates for over 10,000 distinct game titles throughout a socially connected network of nearly 100,000 customers. I augment traditional demand models with descriptive measures of socialization and peer influence. The results suggest that peer effects are strongly influential for demand outcomes after controlling for price changes and product characteristics.

Chapter 3: Recent empirical work suggests that the investment decisions made by household investors are influenced by the opinions of their peers. This paper explores whether social learning can generate beneficial externalities through aggregated crowd wisdom or whether it promotes potentially harmful herding behavior. Using individual investment data from the Motley Fool CAPS platform, I study whether individual predictions can be aggregated to improve the predictability of future asset returns. I provide formal tests of Granger causality and incorporate sentiment indices as covariates into auto-regressive conditional heteroskedasticity (GARCH) models of expected return and volatility. Portfolio simulations demonstrate that, under some circumstances, aggregate sentiment does improve predictability of asset returns in contrast to a more myopic model. Lastly, I quantify the excess portfolio returns provided using a sentiment-aware investment strategy after controlling for common market risk factors.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author wishes to express gratitude to the faculty of the University of Washington Department of Economics for their mentorship, to Skip Sauer, Bobby McCormick, and Mike Maloney who encouraged me to pursue a graduate degree, to my parents Nancy and Don Clayton for their unfaltering support, and to Marie Guldin for her encouragement and motivation.

Lastly, thanks are owed to the peers and faculty who offered valuable advice or feedback throughout this process. Your contributions were highly valued and any remaining errors are my own.

Chapter 1

SOCIAL DETERMINANTS OF ASSET DEMAND: THE IMPACT OF PEER INFORMATION IN ONLINE PREDICTION MARKETS

1.1 Introduction

Household investors face the challenge of composing an investment portfolio by choosing among assets with unpredictable and volatile returns. The prices of these assets, and therefore the returns they provide, change quickly in response to new information or shifts in sentiment. The typical household investor is relatively unsophisticated in his or her ability to predict such changes; frequently turning to the advice of experts or peers to assist with this decision.[1]

Some investors acknowledge their comparative disadvantage, opting for broadly diversified portfolios or market indices which attempt to optimize for return at long horizons. Plentiful evidence also exists for households who unfortunately make serious investing mistakes, either as a consequence of their own lack of sophistication, or by acting upon the poor advice of others. Despite this broad range of expertise, the desire to invest is relatively ubiquitous. As a result, we observe a rich diversity of active participants in the market for household investment, some of whom are largely myopic while others are relatively sophisticated. In many markets, one might expect that such informational asymmetry would stabilize over time, however conditions exist which may also support an equilibrium which maintains this

[1] John Campbell provides a helpful survey which guides the reader through many current challenges in the field of household finance in a 2006 address published in the Journal of Finance.

separation of information.[2]

The heterogeneity amongst stock market participants has generated a meta-market which specializes in the exchange of information between myopic and sophisticated actors, cultivating a large community of professionals and enthusiasts who share investing guidance, ratings of securities, and predictions of future returns. Many popular online communities have developed to specialize in the exchange of investment insights between their own community members.[3] These platforms enable individuals to sample the wisdom of their peers, either individually or in aggregate. One motivation for such sharing of information is a wisdom of crowds hypothesis which suggests that diversity of thought may improve the predictive capabilities of individuals by extending and counter-balancing their own information with the opinions and actions of peers. This paper carefully measures data from one such platform to model the extent to which users are motivated by the choices of their peers and estimates the impact that social interactions have on the investment decisions made by household investors who participate in such communities.

1.1.1 The Motley Fool CAPS Platform

Despite compelling anecdotal and aggregate evidence for the importance of socialization on the household's asset demand, opportunities to examine and quantify these effects at the individual level using real investor data have seldom materialized. An ideal dataset would pair granular individual level demand outcomes with rich measures of socialization between market participants. This data requirement has been largely prohibitive for empirical work,

[2]Gabaix and Laibson (2006) clearly characterize such a "shrouding equilibrium" which may occur in markets where decisions have a high degree of complexity and an asymmetric allocation of information.

[3]In addition to the CAPS platform which this paper studies, notable similar communties include Seeking Alpha, MarketWatch, and TheStreet.

however, the CAPS platform on The Motley Fool presents just such an opportunity.[4] CAPS is a social investment and prediction platform which encourages participants to "become a smarter investor" by using shared tools to rate and research stocks. The CAPS platform accumulates the predictions of its individual users in order to generate aggregate ratings and recommendations based on the collective wisdom of its user-base. CAPS advertises itself with the following tag-line:

> This revolutionary service pools the resources of the Motley Fool Community to help you identify the best stocks at the best times to buy them – and which stocks to avoid.

Of immediate interest to this investigation, the CAPS platform features a compelling combination of pseudo-investment decisions made by thousands of individual users who are linked by an observable network of social connections.[5] The CAPS platform is effectively a game; users (or players) on CAPS issue predictions regarding the future performance of real-world assets and earn a score based on the actual market realizations of their predictions. In this sense, the CAPS game may be aptly characterized as "fantasy football for investing enthusiasts". Players on the CAPS platform interact socially by sharing predictions, discussing investment strategies, and establishing connections with other users.

The primary unit of observation on the CAPS platform is a discrete prediction which indicates whether or not a user expects an individual stock to outperform the S&P 500 index over some optionally specified investment horizon. Participants on CAPS are arguably closer to analysts than investors, issuing asset predictions which are unconstrained by liquidity, ca-

---

[4]Motley Fool (http://www.fool.com) is a financial services company specialized in developing a global investment community. Motley Fool was founded in 1993 and offers a number of core member services including market analysis, an investment newsletter, asset management services, a radio station, a newspaper column, and the CAPS investment platform (http://caps.fool.com). There is no official title for which "CAPS" is an acronym, however platform users have volunteered competing and often amusing hypotheses for the possible meaning of the letters.

[5]I refer to the observed choices on CAPS as pseudo-investments because they are merely predictions, rather than real stock purchases. Players on CAPS share the same incentives that a real investor would, albeit at less risk.

pacity, or other market frictions. These are useful simplifying assumptions, allowing an interpretation of CAPS data as a pure measure of expected return in the absence of many factors which complicate real-world analysis. While the predictions issued by CAPS players do not explicitly map to real-world investment decisions, they are an effective a proxy for how investors behave. CAPS players trade insights in a public forum where reputation is highly valued. The platform assigns each player a public ranking and rewards players who provide effective insights with community prestige through leader-boards which recognize the best performing CAPS players.

In addition to submitting predictions, CAPS users establish "favorite" connections. These are unilateral social relationships between players which provide the initiating user with direct visibility into the predictions issued by the followed player. Player predictions are publicly visible even in the absence of an explicit favorite connection, however the CAPS interface provides users with a more prominent display of the recent activity for their favorite peers.[6] The social tools provided by the CAPS platform allow players to reveal the set of peers whose opinions are of interest to them, a design which facilitates research into how the investment decisions made by household-level investors is influenced by the social group with whom they associate. Figure 1.1 below depicts the observed social network for one player, "vy100", who is one of the most successful players observed in the data.

The player is visualized at the center of the network in green, while his peers are shown in blue. The size and hue of the peer nodes visually denote the number of favorite connections established by those peers. This player has 17 social connections denoted with green lines. Most of these connections are unilateral, depicted with dashed lines, while two are mutual as indicated by solid connectors. This particular network is relatively small as well as highly inter-connected with many mutual connections between peers. We data with these features

[6]The exact interface and functionality presented by the CAPS platform has evolved over time, but not dramatically. This paper assumes that the characteristics of social information conveyed to the user have remained constant over time and that the fundamental relationships between this social information and observed prediction outcomes are, therefore, also invariant.

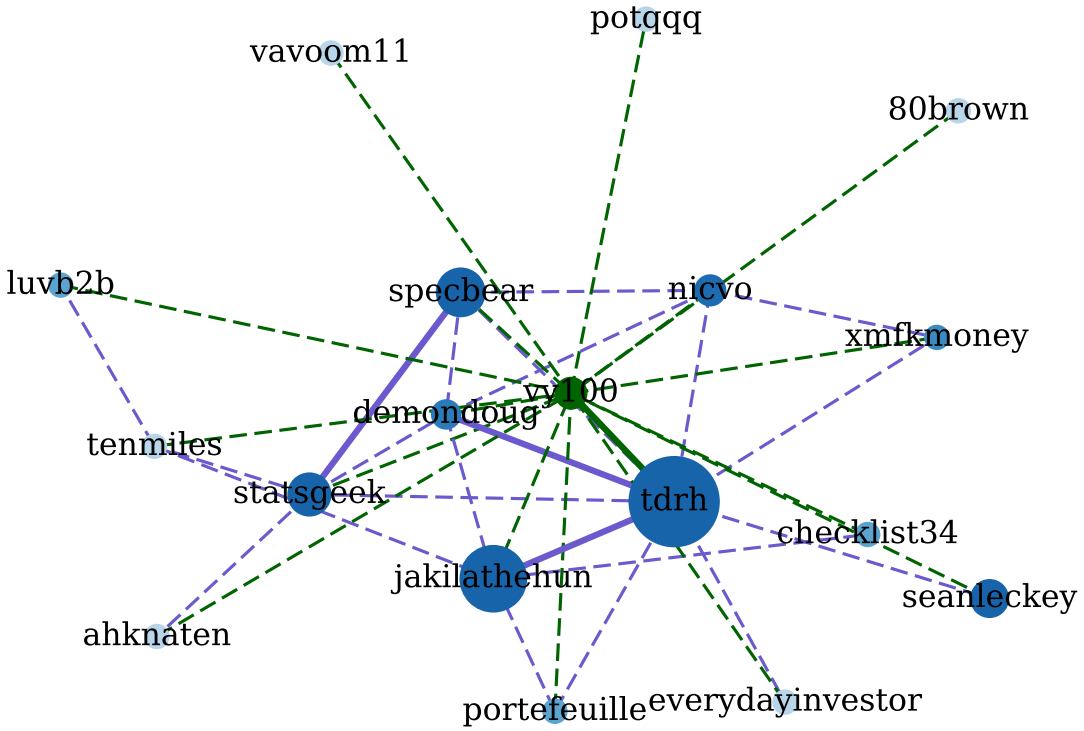

Figure 1.1: CAPS Player Network

for each of the nearly 1,000 users included in this study, the richness and variation in individual networks allows us to effectively model how the predictions issued by a user may influence the decisions made by their peers within the CAPS community.

1.1.2 Related Literature

Research into the influence of social interaction on the behavior of CAPS participants draws from literature on the economics of peer interactions as well as financial models of household-level investment decisions. Household investment belongs to a broad class of problems con-

cerning the consumption of goods which have uncertain characteristics.[7] Consumers in such markets refine their own expectations of value by observing the choices made by peers. These problems are often structured as a game in which participants are heterogeneous in their knowledge or access to data. Players therefore have a range of types from myopic to sophisticated, giving rise to equilibria as described in Gabaix and Laibson (2006) in which myopic users subsidize sophisticated peers in order to avoid making costly mistakes.

Estimation of the causal impact of such interactions on individual choices is challenging. Manski (2000) provides a useful survey of 20th century social interactions literature and the challenges faced by empiricists. Even armed with suitable data, identifying causal relationships attributable to the behavior of peers is difficult due to the "reflection problem" by which individual behavior is endogenously determined by membership in one's peer group.[8] Recent empirical successes in the field have used novel data which can tie individual consumption decisions to measures of socialization. In a study of automobile purchasing by Finnish consumers, Grinblatt et al. (2008) find strong evidence that the purchases of an individual's most proximate peers influence customer purchases of automobiles. Similarly, Sorensen (2006) models the choice of health plan selected by faculty in the University of California system and identifies a clear impact of social information where plan selection is strongly correlated within departments even after controlling for unobservable heterogeneity. Moretti (2011) models how the consumption of movies where quality is uncertain is affected by social learning and the spread of peer information during the weeks following a film's theatrical release.

In the finance literature, there are a number of influential papers that consider the behavioral impact of social information on the choices of individual investors. This literature originates with the seminal paper by Shiller and Pound (1989) who demonstrated using questionnaire

[7]See Barzel (1982) for a broad perspective on the modeling of demand for goods whose true value must be inferred using observable characteristics and the inefficiencies that may arise due to costly measurement.

[8]See Manski (1993) for the original identification and discussion of the "reflection problem" and it's consequences for causal estimation.

data from institutional investors that direct interpersonal communications played a pivotal role in investor decision-making. This phenomenon has been more recently measured by Hong, Kubik, and Stein (2004) who suggest that investors increase their level of stock market participation as more of their peers also participate. They demonstrate that households who attend church or interact frequently with neighbors are substantially more likely to invest than otherwise similar peers who are not socially active. In a later paper, the same authors demonstrate that mutual fund managers are more likely to buy or sell specific stocks if other managers in the same city are also trading that asset. The authors propose an epidemic model where word-of-mouth transfer of information between investors is a significant determinant of market outcomes.[9] Further evidence of the significance of socialization on investment outcomes is contributed by Kaustia and Knüpfer (2012) who study the portfolio returns of geographically-clustered peers and identify that after controlling for market returns, media coverage, and other factors, positive returns within a local cluster encourage stock market entry by non-participating peers. Recently, Burstzyn et al. (2014) conducted a high-stakes field experiment within a financial brokerage that paired brokers and randomized the information each would receive about the other's stock purchase decisions. The authors found statistically and economically significant effects of social learning.

Lastly, there have been two prior econometric analyses using CAPS platform data. Avery, Chevalier, and Zeckhauser (2011) consider the potential of excess returns through collaborative filtering of stock selections by aggregating CAPS player predictions. They find that by leveraging the combined wisdom of the Motley Fool community the potential exists for annual returns up to 9% and that these returns are largely due to stock-picking, particularly for negative predictions where the excess returns are more pronounced. In a similar study, Hill and Ready-Campbell (2011) identify potential for returns in excess of the S&P500 index when constructing prediction portfolios weighted using players CAPS rankings. While both studies effectively examine the aggregate information present within the CAPS platform,

[9]Hong, Kubik, and Stein (2005) examine geographically separated managers and companies to demonstrate this differential likelihood is not attributable solely to local preference or expertise.

there has been no investigation using the full granularity of CAPS platform data into the impact of social connectivity on individual-level asset predictions.

This paper uses a simple framework for describing the CAPS player's decision combined with a rich panel dataset to quantify the effects of socially-transmitted information on the asset predictions registered by CAPS participants. Specifically I address the question of how an investor's likelihood of assuming a position regarding a specific asset changes as that individual observes the decisions made by his or her peers. The primary contribution of this paper to the existing literature is to quantify the importance of socialization in explaining household-level investment decisions by estimating the causal impact of information transmitted between peers on the asset predictions made by CAPS players. Our findings suggest significant causal effects and support hypotheses of social learning. These findings have implications for the design of modern investment platforms and for researchers studying investor behavior; demonstrating the importance of social information in addition to conventionally-studied financial characteristics.

1.2 Structure of the CAPS Game

The objective for each player in the CAPS game is to correctly predict whether assets will outperform or underperform the S&P index. Assets, indexed by j earn a return R^h_{jt} at horizon h from holding the asset from period $t-h$ until period t.[10]

$$R^h_{jt} = \frac{(P_{jt} - P_{jt-1})}{P_{jt-h}} \tag{1.1}$$

Similarly, market returns R_{Mt}, are characterized by the return on the SPY index fund which indexes the S&P500 from period $t-h$ until period t.[11] The player's objective is to identify

[10] Where possible, I use standard notation from John Campbell's textbook "The Econometrics of Financial Markets".

[11] Motley Fool uses the SPY exchange traded fund as the benchmark measurement of S&P500 index returns.

excess return on the market, defined as the difference between asset-specific returns and the S&P market return.

$$Y_{jt} = R_{jt} - R_{Mt} \tag{1.2}$$

Players on the CAPS platform, indexed by i, each possess some asymmetric information θ_{ijt} regarding each stock. Portions of this information may either be common across players or asymmetric, allowing individuals to possess heterogeneous information regarding an asset's potential returns. Using their own private information, players form expectations of forward excess returns which we denote as:

$$E_{ijt}\left[Y_{jt+1}\right] = E\left[Y_{jt+1}|\theta_{ijt}\right] \tag{1.3}$$

Given their expectation of returns, players choose strategies $s_{ijt} \in \{-1, 0, 1\}$ which represent performance predictions corresponding to outperform, neutral, or underperform recommendations respectively. An outperform prediction states that the player expects asset returns to exceed market returns, while an underperform prediction states the player expects asset returns to fall below the market index. Players may also refrain from issuing any prediction. Each prediction earns a score π_{ijt} in the subsequent period equal to the excess return generated by the predicted asset.

$$\pi_{ijt} = s_{ijt-1} Y_{jt} \tag{1.4}$$

Players are rewarded for accurate predictions with community prestige and recognition from a publicly-displayed ranking based on the cumulative score of their predictions.

$$\Pi_{it} = \sum_{d=1}^{t} \sum_{j} \pi_{ijd} \tag{1.5}$$

Each player chooses the set of strategies $S_{it} = \{s_{1t}, \ldots, s_{Jt}\}$ which maximizes the sum of expected utility from forward excess returns conditional on the player's current score and

available information θ_{ijt}.

$$\underset{S_{ijt}}{\text{maximize}} \quad E_{ijt}\left[U\left(\Pi_{it} + \sum_{j} s_{ijt} Y_{jt+1}\right)\right] \tag{1.6}$$

I estimate an empirical solution to this problem which relies on the mapping of observed discrete outcomes onto latent break-points in expected utility.[12] For each candidate strategy s_{ijt} I assume that players operate under the following decision rule.

- ◇ Outperform Prediction: $s_{ijt} = 1$ if $E_{ijt}\left[Y_{jt+1}\right] \geq \bar{\gamma}_{it}$
- ◇ Underperform Prediction: $s_{ijt} = -1$ if $E_{ijt}\left[Y_{jt+1}\right] \leq \underline{\gamma}_{it}$
- ◇ Neutral Prediction: $s_{ijt} = 0$ if $\underline{\gamma}_{it} < E_{ijt}\left[Y_{jt+1}\right] < \bar{\gamma}_{it}$

The parameters $\bar{\gamma}_{it}$ and $\underline{\gamma}_{it}$ reflect time-varying, heterogeneous, and possibly asymmetric risk thresholds. Within this simple model framework, individual users are homogeneous in their preferences and available strategy set, but heterogeneous in their access to information. Differences in the data available to each player is the acting force which produces variation in observed choices. The following section details the unique data sources and features which allow estimation of this model. In section 1.4 I estimate the model by maximum likelihood and present empirical results.

[12] $U(\pi)$ is a Von-Neumann Morgenstern expected utility function. I assume properties of positive marginal utility $U'(\pi) > 0$ for all π and risk aversion $E\left[U(\pi + \epsilon)\right] < U(\pi)$ which are necessary and sufficient conditions to support the observed outcomes of CAPS players issuing neutral predictions for assets with greater expected variance.

1.3 Social Prediction Data

To estimate the effects of socialization on the individual likelihood of asset prediction, I assemble a rich panel dataset constructed from four complementary data sources.[13] The cornerstone data component is the daily prediction history of CAPS players extracted from the Motley Fool website. In addition to prediction histories, we capture the evolution of social networks for each sampled player. Motley Fool has over 1.2 million registered users, most of whom are not active participants on the CAPS platform. In order to overcome technical limitations in data collection it was crucial to apply a sampling strategy which restricted attention to the relevant population of active CAPS participants. I applied a straightforward sampling criteria which restricted the population of Motley Fool users to the sub-population of active CAPS participants by requiring a minimum threshold level of participation within the CAPS game prior to the end of 2012.[14] From this population of eligible users, I randomly selected one-hundred *seed* users to form the basis for the panel and extended the sample to include the additional *peers* with whom a seed user formed a social connection. The combination of seeds and peers generated a social network containing 892 players; 100 seeds and the 792 peers which those seeds designated as favorites.[15] Players appear in the panel gradually over time, entering the dataset on the date of their first observed prediction.

---

[13]CAPS prediction histories and favorite players are publicly visible on each players profile and can be extracted using automated data scraping and HTML parsing techniques. This research utilized JavaScript, but similar results could be achieved using a variety scripting languages like Python, Ruby, or others. Extracting the time-varying history of social connections is more complex, but this history was approximated using a series of monthly snapshots starting in 2010.

[14]Specifically, users must have registered on the CAPS platform prior to January 1, 2012 and issued at least 50 total predictions prior to December 31, 2012 in order to be eligible for selection. There was no requirement on the type of predictions issued or the level of socialization of the user.

[15]The primary reason for limiting sample size in this way was that data needed to be re-sampled periodically in order to update the time-varying social network and backfill historical predictions which were closed or modified. Limiting the sample size to the network generated by 100 seed users provided a compromise between reasonably large sample size and feasibility of routine data collection during the early stages of the project.

The set of predicted assets are matched with historical financial characteristics from the CRSP-Compustat merged database provided by Chicago Booth Business School.[16] Daily adjusted prices and trading volumes for all assets are obtained from the Yahoo Finance API which controls for stock splits, dividends, and other disbursements.[17] Lastly, the panel incorporates daily market-level measurements and market factors obtained from Kenneth French's personal website. These data elements are combined to form a longitudinal panel of players and assets observed each day from 2009 through 2013 where the unit of observation is a prediction issued by a CAPS player regarding a specific asset on a certain day. Each individual observation is augmented with time-varying characteristics which describe the asset, the player, or the overall market factors.

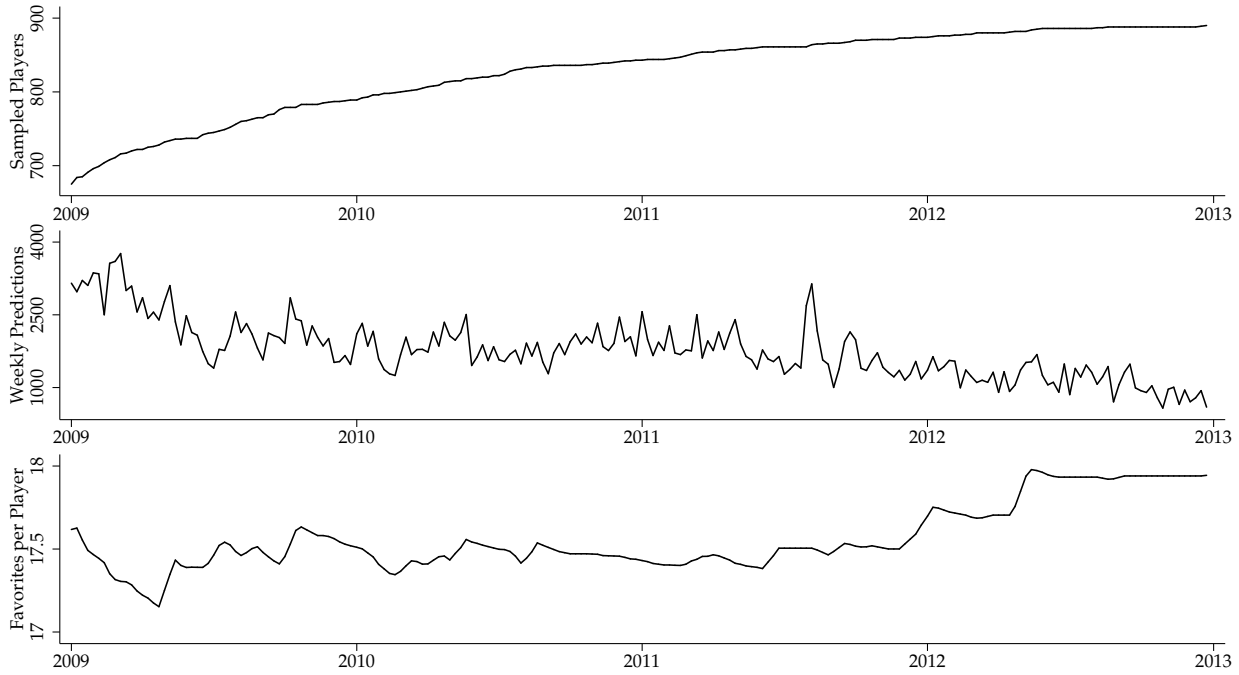

Figure 1.2: Sampled CAPS Activity Over Time

[16]Many thanks to Lewis Thorson in the University of Washington Foster School of Business for his interest in my research and his willingness to help me acquire access to this valuable data source.

[17]Yahoo's ichart provides a method for dynamically retrieving historical price and trade volume data in .csv format using a simple web url-based API.

Figure 1.2 depicts the weekly volume of sampled activity included in the panel. We observe that the pool of sampled users grows steadily throughout the observation period, starting with 650 active players in the beginning of 2009 and growing to 892 players by the end of 2013. New players enter the dataset as they are added as favorite peers of existing seed users. The mean volume of weekly activity has decreased over time. In 2009 the mean active player generated 3.21 predictions per week, decreasing to 1.29 predictions per day by 2013. This phenomenon is due to declining popularity of the CAPS platform as a whole coupled with a tenure effect for observed players; players tend to produce more predictions near the beginning of their time on the platform so as the sampled user segment matures their average level of activity has decreased. We observe cohort effects in both volume of predictions as well as social connections; more recent cohorts of new players tend to establish a larger number of onnections, on average, than cohorts of early platform registrants. As a result, the average number of social connections per player has increased with the mean player increasing from 17.44 favorite users to 17.86 favorites by 2013.

1.3.1 Player Predictions

Our central unit of observation is a single prediction which denotes a position regarding a stock assumed by a player starting on the day the prediction is issued. Each observed prediction is termed a *call* on the CAPS platform and is communicated visually as a thumbs-up or thumbs-down vote. A positive prediction communicates the player's expectation that the asset will outperform the S&P500 index. Players may close predictions at any time, reverting their active position to neutral. Players who do not change or close an active prediction retain their existing position for the subsequent period. Table 1.1 reports the frequency of positions issued by the 100 seed users and their 792 peers by year between 2009 and 2013 during which time 397,505 total predictions were observed.

When submitting a prediction, the player may also optionally include a text-based pitch and designate a suggested term of investment. A pitch describes the rationale motivating

Table 1.1: Player Predictions by Year

Year	Outperform $s_{ijt} = 1$		Underperform $s_{ijt} = -1$		Total
2009	71,399	58.02%	51,655	41.98%	123,054
2010	52,118	54.53%	43,450	45.47%	95,568
2011	41,920	46.58%	48,067	53.42%	89,987
2012	24,401	40.93%	35,212	59.07%	59,613
2013	12,710	43.40%	16,573	56.60%	29,283
Total	202,548	50.95%	194,957	49.05%	397,505

the issued prediction; only 15.26% of submitted predictions include a written pitch. The investment horizon is chosen from a set of options: three weeks, three months, one year, three years, or five years. Most predictions do not designate a specific target horizon. We observe from the data that the median outperform prediction is retained for 54 trading days while the median underperform prediction is held for 42 trading days after it is originally opened.

Throughout the estimation period between 2009 and 2013, users issued predictions for 7,138 distinct recognized assets.[18] Of these recognized assets, 61% are traded on either the New York Stock Exchange (NYSE) or NASDAQ. The remaining 39% are traded either on the American Stock Exchange (ASE), Pacific Exchange (PAC), or through recognized over-the-counter (OTC) exchanges. Each company is also matched to an industry sector classification which describes it's primary domain of operation. Table 1.2 describes the allocation of assets by exchange on which it trades and industry sector in which the firm operates.[19]

[18] An asset is recognized for the purposes of estimation if it is traded on one of the major public exchanges and can be matched by ticker to the CRSP-CompuStat database and the Yahoo Finance API.

[19] We, unfortunately, do not observe an industry sector designation for assets traded on the PAC exchange.

Table 1.2: Predicted Assets by Sector and Exchange

	Traded Exchange					
Market Sector	ASE	NDAQ	NYSE	OTC	PAC	Total
Consumer Discretionary	20	326	284	145	0	775
Consumer Staples	7	81	100	59	0	247
Energy	31	95	270	128	0	524
Financials	14	427	451	133	0	1,025
Health Care	32	436	126	110	0	704
Industrials	19	244	300	117	0	680
Information Technology	19	607	176	117	0	919
Materials	58	51	184	165	0	458
Telecommunications	2	38	44	26	0	110
Utilities	3	14	112	31	0	160
Unknown	115	69	455	10	887	1,536
Total	205	2,319	2,047	1,031	887	7,138

1.3.2 Social Connections

The crucial data feature drawn from the CAPS platform is the social network of connections between users which allows us to model how a player's observed predictions may be influenced by his or her peers. Users in the CAPS network may form unilateral "favorite" connections, through which they receive increased visibility into the predictions made by their followed peer.[20] Within the sampled CAPS data, we observe a network of N nodes, each representing a distinct player i.[21] The state of the CAPS network in period t may be described as a

[20]This type of social connection, used on sites like Twitter, YouTube, and Instagram, is typically referred to as "follow" and generates a directed dynamic graph. This differs from the type of mutual connection typically called a "friendship" which requires the consent of both parties in order to be formed and generates a non-directed dynamic graph. Friendships most notably employed on social sites like Facebook or LinkedIn.

[21]We follow notation proposed in Jackson (2010) for describing such dynamic directed graphs and the meta-data they contain.

directed dynamic graph $(N_t, \boldsymbol{g}_t)$ where $\boldsymbol{g}_t$ is a time-varying adjacency matrix in which an element g_{ikt} denotes whether player i has selected player k as a "favorite" user in or prior to period t. These connections are typically referred to as edges and denoted $ik \in \boldsymbol{g}_t$. For each sampled user $i \in N$, we populate this dynamic adjacency matrix using the approximate dates on which the user formed favorite connections with other players.[22] These social connections describe a set of neighbors for each node i as $N_{it}(\boldsymbol{g}_t) = \{k | ik \in \boldsymbol{g}_t\}$.

The evolving network $(N_t, \boldsymbol{g}_t)$ is comprised of $N = 4,648$ player nodes of whom 100 are seed users, 792 are neighbors of those seeds, and the remaining 3,756 are second-order neighbors or "friends-of-friends." With edges ik in a directed network we are able to categorize player k as a sender of information where player i is a receiver. These directed edges in addition to the time variation in the dynamic network graph provide a mechanism for identification of causal parameters on the influence of peer predictions on the portfolio choices made by user i. Furthermore, for each pair of players $ik \in \boldsymbol{g}_t$, we can categorize one player as a leader while the other is a follower based on the relative ranking of the two players as given by their accumulated scores Π_{it} and Π_{kt}. As a result, the observed asset predictions communicated between players on the CAPS platform are classified along three dimensions: signal (positive or negative), direction (sender or receiver), and standing (leader or follower).

Figure 1.3 visualizes the network of seeds and peers at the end of the observation period in December 2013. The size of nodes reflects the number of neighbors for each user, larger nodes denote users with more favorite connections. Nodes shaded in green are the original seeds, while nodes shaded in blue are their peers. The algorithm which produces this figure clusters points that share common connections, therefore the outer-most nodes in the graph generally represent users who are followed by others but do not, themselves, follow other users within the sample. The inner-most clusters contain groups of highly-connected users who share many common peers. The median user has 12 neighbors while the most socialized

[22] These true connection creation dates are not directly observable on the CAPS platform, but can be inferred with reasonable precision from the sequence of monthly snapshots collected over the course of the sampling period.

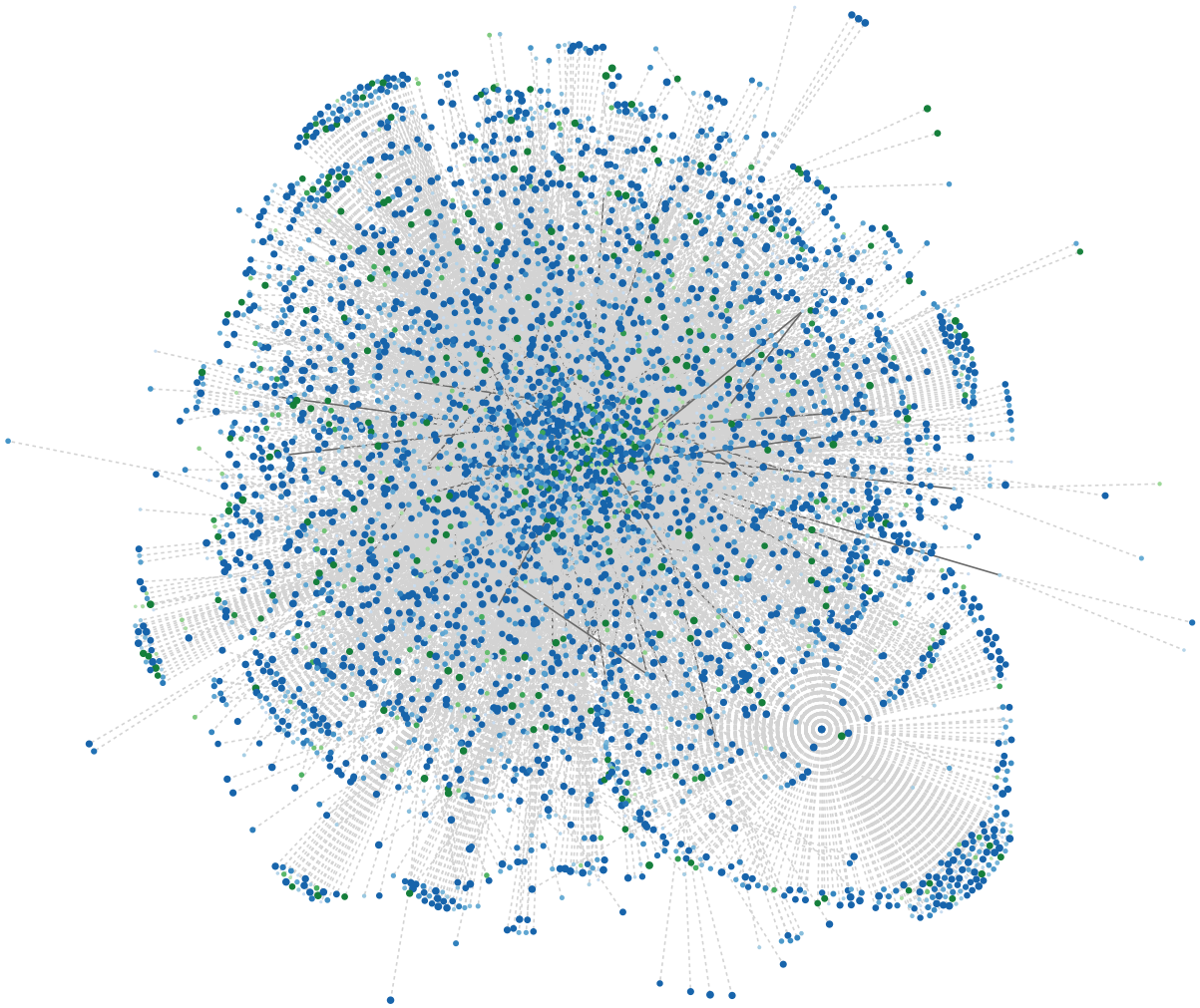

Figure 1.3: Sampled Social Network - December 2013

players within the sample have over 100 connections. This visualization reveals the high degree of connectivity within the dataset. A connection between users depicted as a light, dotted line is a unilateral connection. A solid, dark line represents a bilateral friendship. An expanded plot of the complete network which includes acquaintances is available in Appendix .3.

1.3.3 Estimation Panel Structure

The observational data collected from the CAPS platform begins as the history of non-neutral chosen strategies $s_{ijt} \in \{-1, 1\}$. In order to estimate the marginal effect of demand determinants I expand and transform the data into a dynamically-balanced panel of individuals $i \in N_t$ and assets $j \in J_t$. Individual users enter the panel starting on the date of their first observed prediction.[23] Assets are included in the panel if they are able to be matched with reported financial characteristics and price histories in addition to a materiality threshold which filters for assets which have received any prediction by an observed user over a rolling four-week period.[24] Within a single trading day the panel is strongly balanced such that active players N_t may issue predictions over a common set of assets J_t. Over time, the composition of the daily cross-section changes as additional players and different assets enter the data.

The resulting panel features daily data from January 2009 through December 2013, a period which spans the height of CAPS platform popularity. During this period 6,163 distinct assets satisfy the criteria inclusion in J_t for at least some period t. Of these active assets, 427 are omnipresent throughout all four years of the sample. Out of the 100 seed users, 64 had begun activity on the CAPS platform prior to the beginning of 2009 while the remaining 36 enter the panel following the date of their first observed prediction on the CAPS platform. The complete dataset containing the history of prediction decisions made by these 100 seed users over the set of active securities contains 85,079,589 observations, where each observation describes a user's prediction opportunity regarding a certain asset on a specific trading day. The dependent variable in our model, s_{ijt} is a categorical outcome variable reporting the strategy chosen in each prediction opportunity. The frequencies of observed prediction outcomes are reported in Table 1.3.

[23] We don't observe the true registration date for each CAPS user, so it is only possible to add them to the set of active players on the date of their first observed activity.

[24] There are several other minor restrictions which eliminate stocks from consideration. For example I exclude predictions for assets which have not yet gone public.

Table 1.3: Dependent Variable - Chosen Strategy s_{ijt}

No Prediction	Outperform	Underperform
$s_{ijt} = 0$	$s_{ijt} = 1$	$s_{ijt} = -1$
84,995,238	61,156	73,748
99.901%	0.072%	0.087%

Frequency of outcomes reported for total sample size $n = 85,079,589$.

Non-neutral predictions are rare given this data structure which can be a significant concern under estimation by maximum likelihood due to the infrequence of outcomes of interest. King and Zeng (2001) show that bias introduced due to rare outcomes is proportional to the number of outcomes in the least-frequently observed category. In this case, despite the rarity of events, the relatively large number of observed outperform predictions reasonably remediates this concern.[25]

1.4 Estimation Strategy

Using the latent utility model framework proposed in section 2.3, we approximate player i's unobserved expected utility of issuing a prediction, $E_{ijt}\left[U(Y_{ijt})\right]$, as a linear combination of observable features. These explanatory variables are conveniently categorized using four thematic groupings. Financial fundamentals F_{jt} vary by asset and day and include historical returns, trading volumes, features from quarterly revenue reporting, and industry categorization. Macroeconomic factors M_t include features of market-level returns and time fixed effects. Investor heterogeneity I_{it} captures characteristics which vary by individual user including measures of tenure, historical performance, and recent activity. Lastly, social information S_{ijt} captures the private information observed by each user through their own

[25]See King and Zeng (2001) for exact detail of the asymptotic properties and consistency of maximum likelihood type estimators when performing inference using rare events data. The authors show that the risk of bias is proportional to the number of observations in the least-frequently observed category.

heterogeneous and dynamic social network $(N_{it}, \boldsymbol{g}_t)$.

A full definition of covariate features included in the model is provided in Appendix .4 on page 100 and these features are summarized in Appendix .2. For convenience in notation, I denote the expected utility from selecting strategy $s_{ijt} = c$ for $c \in \{-1, 1\}$ as y^c_{ijt} which is modeled as:

$$\begin{aligned} y^c_{ijt} &= E_{ijt}\left[U(s_{ijt}Y_{jt+1})|s_{ijt} = c\right] \\ &= \alpha^c + \beta^c_F F_{jt} + \beta^c_M M_t + \beta^c_I I_{it} + \beta^c_S S_{ijt} \end{aligned} \tag{1.7}$$

We estimate the probability of observing each type of prediction using the multinomial logistic framework where the baseline outcome is no prediction $s_{ijt} = 0$. For each prediction opportunity, y_{ijt}, the included covariates capture information available to player i prior to the beginning of period t in order to protect against issues of endogeneity due to unobserved intra-day timing or reflection. I estimate the probability for $s \in \{-1, 1\}$ of observing each prediction where the parameters are estimated by maximum likelihood and express the change in log-odds of issuing a certain prediction relative to the baseline outcome.[26]

A critical assumption of the multinomial logistic model is that it satisfies the axiom of independence of irrelevant alternatives (IIA) which requires that if an alternative x is chosen from a set of alternatives T, and x is also an element of a subset S of T, then x must be the alternative chosen from S. As demonstrated by McFadden (1974), this assumption can be easily violated in many applications if the inclusion of additional alternatives in the option set dilutes the distinct nature of each option. In the presence of such irrelevant alternatives, estimating the likelihood of observing an outcome using the multinomial logistic model is invalid since individual preference between alternatives depends on the alternative set which is offered. In McFadden's own words, "application of the [multinomial logistic] model should

[26] An alternative model specification would be to rely upon an implicit ordering of outcomes, where negative predictions imply the lowest expected returns relative to the S&P index and positive predictions imply the highest expected returns. While observed predictions do have properties of ordered responses, an ordinal logistic model it not appropriate for estimation in this context due to (1) a violation of the proportional odds assumption and (2) extreme over-representation of the middle-category relative to the two outcomes of interest which would bias the estimated marginal effects.

be limited to situations where the alternatives can plausibly be assumed to be distinct and weighed independently in the eyes of each decision-maker". In the case of Motley Fool data, the three possible predictions constitute the unrestricted option set from which no element can be removed which would cause this axiom to fail. If a CAPS player issues an outperform prediction for a certain asset, that player would still select the same strategy even if the option of registering an underperform or neutral prediction were removed.

Another assumption of the multinomial logistic structure is that the data is case specific. In other words, the value of included covariates are the same for each available strategy. This assumption is upheld in the structure of included covariates. Players issue directional predictions given an the information set θ_{ijt} which varies by asset and over time, but not by strategy. Therefore, the prerequisite assumptions for appropriate use of the multinomial logistic model are satisfied under the structure of the CAPS game.

1.4.1 Primary Hypotheses

This empirical framework allows for testing of several hypotheses regarding the responsiveness of investor behavior to the observable choices of their peers. The first hypothesis upon which others are predicated is a *social information hypothesis* (H_0^1) which postulates that after controlling for publicly observable asset characteristics the behavior of peers is a causal predictor of the investment choices selected by a socially networked individual. I test that the inclusion of heterogeneous social covariates S_{ijt} meaningfully improves the explanatory power of our estimated model as measured by the log-likelihood of observed asset demand under both restricted and unrestricted model specifications.

A related investigation is an exploration of Hong, Kubik, and Stein's (2004) *market participation hypothesis* (H_0^2) which examines whether increases in levels of socialization contribute directly to a increase in individual likelihood to invest. In the CAPS context, I test this hypothesis by evaluating whether we observe a measurable increase in the likelihood of

observing outcomes of interest $s_{ijt} \in \{-1, 1\}$ for users who possess a greater number of neighbors N_{it} or whose neighbors generate a greater volume of observed market activity.

An additional hypothesis which has narrowly focused implications for the design of CAPS or other similar investment products involves whether investors behave strategically in the way they respond to different flavors of social information. I propose a *strategic response hypothesis* (H_0^3) which suggests that users who observe predictions originating from more highly ranked peers may attribute greater weight towards those predictions while discounting those generated by lower ranked players. The directed graph structure of the CAPS platform allows for us to identify possibly asymmetric effects in the efficacy of peer influence. I test this hypothesis by evaluating the restriction that the predictions of "leaders" and "followers" are equally influential in their causal impact on demand.

Lastly, this framework provides a suitable platform for quantifying the impact of financial covariates which are typically included in models of household-level investment and contrast their influence on individual demand with the more nuanced measures of social information which the CAPS dataset provides. In the following section, I present empirical results and test these hypotheses.

1.5 Econometric Results

I estimate two specifications of the multinomial logistic model described in section 1.4. The Baseline Model specification excludes social covariates S_{ijt} from the estimation equation and only includes financial fundamentals F_{jt}, macroeconomic factors M_t, and investor heterogeneity I_{it}. The Social Model specification augments the estimation equation with these social covariates which vary by player, asset, and time. Results from both models are reported in table 1.4 on pages 23 and 24. The interpretation of the reported coefficients is the marginal effect of a one-unit change in the covariate on the log-odds of issuing an outperform or underperform prediction relative to the base case of issuing no prediction for that asset.

Table 1.4: Multinomial Logit Results

Covariate	Baseline Model		Social Model	
	Outperform	Underperform	Outperform	Underperform
$volume_{jt}$	0.0034*	0.0030*	0.0030*	0.0027*
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
$alpha_{jt}$	0.0154	-0.1007*	0.0011	-0.0747*
	(0.0225)	(0.0230)	0.0210)	(0.0253)
$beta_{jt}$	-0.3907*	-0.6298*	-0.2998*	-0.5526*
	(0.0501)	(0.0426)	(0.0482)	(0.0506)
$prior\_return_{jt}$	-0.0043	-0.0174*	-0.0020	-0.0131*
	(0.0032)	(0.0037)	(0.0029)	(0.0039)
$market\_return_{t}$	-0.0344	-0.0268	-0.0291	-0.0227
	(0.0290)	(0.0196)	(0.0237)	(0.0192)
$small\_minus\_big_{t}$	0.0318	0.0210	0.0251	0.0134
	(0.0368)	(0.0194)	(0.0381)	(0.0196)
$high\_minus\_low_{t}$	0.0659	0.0696*	0.0516	0.0644*
	(0.0359)	(0.0303)	(0.0412	(0.0305)
$momentum_{t}$	-0.0007	-0.0423*	0.0019	-0.0357
	(0.0286)	(0.0192)	(0.0286)	(0.0191)
$lifetime\_picks_{it}$	0.0006*	0.0007*	0.0004*	0.0004*
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$player\_tenure_{it}$	-0.0539*	-0.0341*	-0.0487*	-0.0264*
	(0.0057)	(0.0056)	(0.0059)	(0.0060)
$player\_pick\_up_{it}$	0.0206*	0.0079	0.0198*	0.0071
	(0.0029)	(0.0044)	(0.0028)	(0.0046)
$player\_pick\_down_{it}$	0.0028	0.0155*	0.0020	0.0144*
	(0.0031)	(0.0031)	(0.0032)	(0.0029)
$player_i$ fixed effects	✓	✓	✓	✓
$exchange_j$ fixed effects	✓	✓	✓	✓
$sector_j$ fixed effects	✓	✓	✓	✓

This table is continued on the following page.

Estimates marked (*) are statistically significant at the 1% level. Standard errors are reported in parentheses and are clustered by player i.

Variable subscripts describe their level of variation in the data. Trading days are indexed by t, players are indexed by i, and assets are indexed by j.

Table 1.4: Multinomial Logit Results (Continued)

Covariate	Baseline Model		Social Model	
	Outperform	Underperform	Outperform	Underperform
$total\_favorites_{it}$	-	-	0.0315*	0.0620*
			(0.0114)	(0.0181)
$leader\_pick\_up_{ijt}$	-	-	0.4659*	0.3742*
			(0.1175)	(0.0848)
$leader\_pick\_down_{ijt}$	-	-	0.2717*	0.4428*
			(0.0927)	(0.1188)
$follower\_pick\_up_{ijt}$	-	-	0.3432*	0.6308*
			(0.0760)	(0.0932)
$follower\_pick\_down_{ijt}$	-	-	0.4732*	0.2108*
			(0.1118)	(0.1106)
$leader\_hold\_up_{ijt}$	-	-	1.3238*	1.1516*
			(0.1967)	(0.1231)
$leader\_hold\_down_{ijt}$	-	-	1.1272*	1.4059*
			(0.1351)	(0.1510)
$follower\_hold\_up_{ijt}$	-	-	1.1552*	1.3667*
			(0.1281)	(0.1332)
$follower\_hold\_down_{ijt}$	-	-	1.4781*	1.1806*
			(0.1217)	(0.1568)
McFadden R^2	0.1257		0.1515	
Observations	85,079,589		85,079,589	

This table is continued from the previous page.
Estimates marked (*) are statistically significant at the 1% level. Standard errors are reported in parentheses and are clustered by player i.
Variable subscripts describe their level of variation in the data. Trading days are indexed by t, players are indexed by i, and assets are indexed by j.

I control for financial fundamentals by including daily trading $volume_{jt}$, time-varying estimates of long-run $alpha_{jt}$ and $beta_{jt}$ as defined by the traditional CAPM model, trailing 7-day $prior\_return_{jt}$, and fixed effects for the industrial sector in which the firm produces and exchange on which the asset trades. These are publicly available financial characteristics which even relatively unsophisticated investors might consider when issuing predictions.

Time-varying macroeconomic factors include prior-week market returns. Differences in the

composition of market returns are reported by the Fama-French factors $small\_minus\_big_t$ which tracks the excess returns from small capitalization securities over large capitalization stocks and $high\_minus\_low_t$ which measures excess returns of value stocks (high book-to-market ratio) compared to growth stocks (low book-to-market ratio).[27] We also include Carhart's momentum factor which captures the market's tendency to continue moving in the same direction over the past 30 trading days.[28]

I control for some degree of individual heterogeneity through the historical predictions made by sampled players. Each player's cumulative prediction activity is tracked by $lifetime\_picks_{it}$ which reports the total number of predictions issued by that player to-date. Tenure on the CAPS platform, $player\_tenure_{it}$, measures the number of days since the user first participated on the CAPS platform after registering an account. The variables $player\_pick\_up_{it}$ and $player\_pick\_down_{it}$ report the number of directional predictions issued by that user within the past week. Additional unobserved heterogeneity is controlled for using a player-specific fixed effect.

The core set of social covariates measure the level of socialization and transmission of social information between peers. For each player, $total\_favorites_{it}$ counts the number of other users which the player has marked as favorites. I also include a set of social variables which track the number of new and active predictions made for each direction (up or down) and standing (leader or follower) tuple. Therefore, $leader\_pick\_up_{ijt}$ and $follower\_pick\_up_{ijt}$ report the number of positive predictions regarding asset j originating from favorites of player i who are more (or less) highly ranked in period t. Similarly $leader\_holds\_down_{ijt}$ and $follower\_holds\_down_{ijt}$ report the number of player i's favorites who hold an active underperform position regarding asset j in period t.

---

[27] Fama and French, *Common risk factors in the returns on stocks and bonds.* 1992. Data obtained from Kenneth Frenchs personal webpage: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french

[28] Carhart, Mark M. *On Persistence in Mutual Fund Performance*, 1997.

1.5.1 The Influence of Social Information

These results clearly emphasize the influence of social effects on the observed asset selection decisions of participants in the CAPS platform. The estimated coefficients reported for non-social covariates on page 23 are stable between the baseline and social model specifications, suggesting that the inclusion of peer information adds an incremental layer of explanatory power rather than absorbing information already contained in the included non-social covariates. To test the proposed *social information hypothesis* (H_0^1), we can contrast the likelihood of explaining the observed predictions. A comparison of the pseudo-R^2 statistics suggests that compared to the baseline specification, inclusion of social information increases the explanatory power of the model by 20%.[29]

$$\begin{aligned} &H_0^1 : \mathcal{L}^{\text{Baseline}} = \mathcal{L}^{\text{Social}} \\ &LR = -2\left[\log(\mathcal{L}^{\text{Baseline}}) - \log(\mathcal{L}^{\text{Social}})\right] \sim \chi^2(18) \end{aligned} \tag{1.8}$$

A likelihood-ratio test that the constrained baseline model captures the same information as the unconstrained social model produces a test statistic of $LR = 10,966.21$ which overwhelmingly rejects the null hypothesis under the $\chi^2(18)$ distribution.[30] This finding suggests compelling evidence that the inclusion of social information is powerful in explaining the behavior of investors within the CAPS context while possibly contributing similar value to more generalized models of household investor behavior.

Coefficient estimates which report the marginal effects on the empirical log-odds ratios are presented for social covariates in the second half of table 1.4 on page 24. I estimate coefficients for each of the nine social covariates which are large in both magnitude and statistical significance. The coefficients on $total\_favorites_{it}$ suggest that the individual likelihood of

[29] McFadden's pseudo R^2 is computed as $R^2 = 1 - ln\left[L(M^{full})\right] / ln\left[L(M^{intercept})\right]$ where L is the models estimated likelihood score. It can be interpreted as a measure of the gains in informational content provided by the full model over an intercept-only specification.

[30] The LR test statistic represents the information gained from the unconstrained social model relative to the baseline non-social model.

issuing a non-neutral prediction increases by roughly 3.2% for positive predictions and 6.2% for negative predictions respectively for each additional social connection established by the player, after controlling for other factors. This finding provides preliminary support for the *market participation hypothesis* (H_0^2) which speculates that an investor's level of market participation increases with his or her level of socialization with other investors. We can further consider how the observed level of market participation changes as a user's peers also hold a position regarding the same asset.

The coefficient on $leader\_hold\_up_{ijt}$ suggests that each additional higher-ranked favorite who holds a positive position on a specific stock increases the user's likelihood of also taking a positive position by 132% and the likelihood of instead taking a negative position by 115%. We observe similar large effects for followers and for peers who hold an underperform position. I formally test the market participation hypothesis by evaluating the joint significance of the social variables which denote whether peers also hold a specific asset. We can define the joint hypothesis test as follows where β_{hold} denotes the set of coefficients corresponding to social covariates measuring how many peers currently hold active positions.

$$\begin{aligned} &\beta^o_{hold} = \beta^o_{leader\_hold\_up}, \beta^o_{leader\_hold\_down}, \beta^o_{follower\_hold\_up}, \beta^o_{follower\_hold\_down} \\ &H_0^2 : \sum \beta^o_{hold} = 0 \quad \text{where} \quad F^o_{hold} \sim \chi^2(2) \end{aligned} \tag{1.9}$$

A joint F-test using this linear restriction produces test statistics of $F^o_{hold} = 148.29$ for outperform predictions and $F^u_{hold} = 139.16$ for underperform predictions against a $\chi^2(2)$ distribution. Both tests reject the null hypothesis that the level of peer engagement in the market for an asset has no influence on an individual's likelihood to participate in the same market.

The third hypothesis of interest to our investigation is whether users behave strategically, contextualizing their response to observed peer predictions based on the relative platform standing of that peer. Our hypothesis is that players are more likely to match the directionality of an observed prediction if it originates from a more sophisticated peer than if

the issuing peer is less highly ranked. To test this *strategic response hypothesis* (H_0^3) I contrast the response taken to new predictions observed within the previous 7 days. I test this hypothesis against the linear restriction that players are equally likely to match a peer's prediction regardless of whether that peer is a leader or a follower.

$$H_0^3 : \underbrace{\beta_{leader\_up}^{outperform} + \beta_{leader\_down}^{underperform}}_{\text{Imitate Leader}} = \underbrace{\beta_{follower\_up}^{outperform} + \beta_{follower\_down}^{underperform}}_{\text{Imitate Follower}} \tag{1.10}$$

Furthermore we can contrast whether, conditional on observing a prediction issued by a more sophisticated leader, the player is more likely to imitate or deviate from that prediction.

$$H_0^3 : \underbrace{\beta_{leader\_up}^{outperform} + \beta_{leader\_down}^{underperform}}_{\text{Imitate Leader}} = \underbrace{\beta_{leader\_up}^{underperform} + \beta_{leader\_down}^{outperform}}_{\text{Deviate from Leader}} \tag{1.11}$$

Table 1.5: Strategic Comparison - Cumulative Relative Risk

	Imitate	Deviate	F-stat
Leader	3.15	2.77	2.53
Follower	2.64	3.48	25.50*
F-Stat	2.87*	22.92*	-

F-statistics are tested against a $\chi^2(1)$ distribution.
Statistics marked (*) are significant at the 10% level.

Table 1.5 reports cumulative relative risk ratios of imitating versus deviating in response to an observed peer prediction. We observe moderate evidence based on the estimated difference in cumulative risk ratios that CAPS players are more likely to imitate a leader than a follower. More strikingly, players are much more likely to deviate from a follower than a leader. Comparing responses to signals of the same type reveals that there is no strong statistical difference in the estimated response to a leader's prediction, while there is strong evidence that players are more likely to deviate from followers than imitate them. Overall, these results suggest compelling evidence of strategic behavior. This motivates a more broad

research question as to whether household investors tend to qualitatively classify their peers are more or less sophisticated in their investing acumen.

1.5.2 Evidence from Non-Social Covariates

The non-social covariates included in our model, while not the primary focus of this paper, do also demonstrate some interesting features. We observe that players rationally respond to historically low rates of alpha by increasing their likelihood of predicting underperformance. Similarly, we observe players decreasing their likelihood of issuing a prediction for assets which have higher beta and therefore greater covariance with the S&P500 index. Additionally, players are more likely to predict underperformance immediately following a week of poor performance. These findings suggest that participants on the CAPS platform are backwards looking, at least to some degree, using historical variation in asset returns to inform their forward looking predictions. In contrast, the log-odds estimates for macroeconomic factors show that CAPS players exhibit almost no response to market-level return factors with the possible exception of HML, which measures the spread between value and growth stocks.[31]

We observe that individual characteristics can explain meaningful variation in prediction outcomes. CAPS players are unsurprisingly more likely to issue a prediction (either positive or negative) if they have generated a larger volume of past predictions, however this effect is offset by a negative effect of tenure. For each additional month that a player participates on CAPS, their log-odds of issuing additional predictions decreases by between 2.5 and 5%, suggesting that the novelty of participating in the CAPS prediction game may wear off over time. A player's recent sentiment captures a meaningful difference in their local likelihood of issuing new predictions. For each additional outperform prediction issued in the trailing 7 days, the average player is 2% more likely to issue a positive prediction than a negative

[31] *Value* stocks are defined as having a large book-to-market ratio, while *growth* stocks have a low ratio.

one. Similarly, for each recent underperform prediction, the typical user is 1.6% more likely to issue a negative prediction. This suggests that the sentimentality of individual investors is material, after controlling for other factors, in explaining the positions which they choose to select.

1.5.3 Robustness Against Selection Bias

A reasonable concern regarding these estimates is the risk of selection bias if CAPS players systematically select favorites whose predictions tend to reinforce their own investment preferences. In this way, certain CAPS players may be inherently more likely to act upon observed social information, possibly introducing upward bias into the estimated impact of socially-transmitted information. To understand the extent of this concern, I modify the model presented in Section 1.5 by implementing a first-stage propensity score estimator which stratifies players into time-varying propensity groups and evaluates the stability of model estimates across groups.[32] To determine group assignment, I model an index of socialization which corresponds to the volume of social information each player receives from their peer network. Specifically, where $N_{it}(\boldsymbol{g}_t)$ denotes the neighbors for player i:

$$socialization_{it} = \sum_{j} \sum_{k \in N_{it}} |s_{kjt}| \tag{1.12}$$

We predict the period $t+1$ level of socialization as a function of individual and market characteristics.

$$socialization_{it+1} = \alpha_i + \phi \log(trend_t) + \beta_M M_t + \beta_I I_{it} + \epsilon_{it} \tag{1.13}$$

The predicted level of social engagement is used to stratify players into groups where higher indexed groups have a greater expected propensity for social engagement. Classifying this

[32]The textbook *Causal Inference for Statistics, Social, and Biomedical Sciences* by Imbens and Rubin provides an excellent resource on assignment mechanisms, particularly propensity score stratification.

expected level of socialization is a pure prediction problem, so I obtain these estimates using a gradient-boosted regression tree. The matrices M_t and I_{it} correspond conceptually to the macroeconomic factors and individual heterogeneity controlled for in the primary model, but include a more rich set of covariates interaction terms. Evidence that classification or regression tree methods can improve covariate selection and reduce bias in propensity score matching is documented in Diamond and Sekhon (2013). The gradient-boosted regression tree is trained using data from 2009 and 2013 and then predicted for data in the primary sample period of 2010 through 2012. The out-of-sample predicted values are used as estimates of propensity for each individual to be exposed to social information.

Upon retrieving the predicted propensities from the random forest model, I stratify the overall sample into ten bins using deciles of this propensity score. The multinomial logistic model from Section 1.5 is estimated within each propensity bin. The resulting coefficients and variances are weighted accross bins for each covariate as shown below. Weighted cross-bin estimates of multinomial logic coefficients and standard errors are reported in table 1.6 on pages 32 and 33.

$$\hat{\beta} = \sum_{g=1}^{10} \left(\frac{N_g}{N} \right) \hat{\beta}_g \qquad \hat{\sigma}^2 = \sum_{g=1}^{10} \left(\frac{N_g}{N} \right) \hat{\sigma_g^2} \tag{1.14}$$

These coefficient estimates on social covariates are similar in magnitude to those produced by the original model, providing some reassurance that our initial estimates are not overly affected by upwards bias due to endogenous selection. Under the stratified multinomial framework, coefficients are attributed higher measures of error due to high variance of the within-bin estimates particularly for lower indexed deciles of socialization. In these lower bins the probability of observing a prediction conditional on a non-zero volume of social signals is very low. After re-weighting estimated log-odds and standard errors across bins the overall intuition regarding the influence of socialization on predictive outcomes is preserved, suggesting these estimates capture a meaningful behavioral phenomenon rather than the results of selection based on observables.

Table 1.6: Propensity Binned Results

	Standard Model		Binned Model	
Covariate	Outperform	Underperform	Outperform	Underperform
$volume_{jt}$	0.0030* (0.0002)	0.0027* (0.0002)	0.0024* (0.0004)	0.0020* (0.0004)
$alpha_{jt}$	0.0011 0.0210)	-0.0747* (0.0253)	0.0365 (0.0333)	-0.0931* (0.0290)
$beta_{jt}$	-0.2998* (0.0482)	-0.5526* (0.0506)	-0.2990* (0.0630)	-0.4797* (0.0701)
$prior\_return_{jt}$	-0.0020 (0.0029)	-0.0131* (0.0039)	-0.0035 (0.0067)	-0.0130* (0.0061)
$market\_return_{t}$	-0.0291 (0.0237)	-0.0227 (0.0192)	-0.0418 (0.0382)	-0.0101 (0.0338)
$small\_minus\_big_{t}$	0.0251 (0.0381)	0.0134 (0.0196)	0.0389 (0.0671)	0.0119 (0.0428)
$high\_minus\_low_{t}$	0.0516 (0.0412	0.0644* (0.0305)	0.0729 (0.0739)	0.0440 (0.0615)
$momentum_{t}$	0.0019 (0.0286)	-0.0357 (0.0191)	-0.0125 (0.0518)	-0.0354 (0.0330)
$lifetime\_picks_{it}$	0.0004* (0.0001)	0.0004* (0.0001)	0.0001 (0.0002)	0.0001 (0.0003)
$player\_tenure_{it}$	-0.0487* (0.0059)	-0.0264* (0.0060)	-0.0271* (0.0116)	-0.0099 (0.0102)
$player\_pick\_up_{it}$	0.0198* (0.0028)	0.0071 (0.0046)	0.0159 (0.0144)	0.0015 (0.0107)
$player\_pick\_down_{it}$	0.0020 (0.0032)	0.0144* (0.0029)	-0.0011 (0.0074)	0.0106* (0.0042)
$player_i$ fixed effects	✓	✓	✓	✓
$exchange_j$ fixed effects	✓	✓	✓	✓
$sector_j$ fixed effects	✓	✓	✓	✓

This table is continued on the following page.

Estimates marked (*) are statistically significant at the 1% level. Standard errors are reported in parentheses and are clustered by player i.

Variable subscripts describe their level of variation in the data. Trading days are indexed by t, players are indexed by i, and assets are indexed by j.

Table 1.6: Propensity Binned Results (Continued)

Covariate	Standard Model		Binned Model	
	Outperform	Underperform	Outperform	Underperform
$total\_favorites_{it}$	0.0315*	0.0620*	0.0786*	0.0794*
	(0.0114)	(0.0181)	(0.0380)	(0.0326)
$leader\_pick\_up_{ijt}$	0.4659*	0.3742*	0.9217*	0.5893*
	(0.1175)	(0.0848)	(0.3822)	(0.2318)
$leader\_pick\_down_{ijt}$	0.2717*	0.4428*	0.3015	0.9165*
	(0.0927)	(0.1188)	(0.2914)	(0.2605)
$follower\_pick\_up_{ijt}$	0.3432*	0.6308*	0.4455	0.7609
	(0.0760)	(0.0932)	(0.2306)	(0.4326)
$follower\_pick\_down_{ijt}$	0.4732*	0.2108*	0.5224*	0.2528
	(0.1118)	(0.1106)	(0.2058)	(0.1562)
$leader\_hold\_up_{ijt}$	1.3238*	1.1516*	1.4517*	0.7360*
	(0.1967)	(0.1231)	(0.2225)	(0.2027)
$leader\_hold\_down_{ijt}$	1.1272*	1.4059*	0.9961*	2.1007*
	(0.1351)	(0.1510)	(0.2987)	(0.4877)
$follower\_hold\_up_{ijt}$	1.1552*	1.3667*	0.5787	1.0013
	(0.1281)	(0.1332)	(0.7410)	(1.1211)
$follower\_hold\_down_{ijt}$	1.4781*	1.1806*	1.9149*	0.9263*
	(0.1217)	(0.1568)	(0.4436)	(0.2317)
Propensity Bins	-		10	
Observations	85,079,589		85,079,589	

Table continued from the previous page.

Estimates marked (*) are significant at the 1% level. Standard errors are clustered by player i. Variable subscripts describe their level of variation in the data. Trading days are indexed by t, players are indexed by i, and securities are indexed by j. All variables capture information available at the beginning of trading day t.

1.6 Conclusion and Future Work

This paper contributes to a growing body of work that empirically quantifies the importance of peer effects within markets which feature a social component. Specifically within the field of household finance, where the distribution of information across market participants may be highly asymmetric, these findings reinforce evidence that individuals consider the invest-

ment decisions of their peers when making asset selection decisions. Our findings show that visibility into the decisions made by peers almost certainly has the ability to influence the choices made by individual investors. Furthermore, within the context of the CAPS platform, I show strong strategic effects where individuals qualitatively weight the sophistication of their peers when evaluating their choices. There is considerable opportunity to perform further analysis both using CAPS asset prediction data; a second paper investigates the extent to which the aggregated social wisdom of CAPS participants can contribute meaningfully to a portfolio selection strategy by testing the influence of aggregate indices of social information on market outcomes.

While the structure of the CAPS platform provides an attractive sandbox for evaluating the way that investing enthusiasts form expectations of asset returns, there is much greater potential for continuing this work using actual household-level investment data. The results of this paper are strongly suggestive of the high value individual investors place on observing the financial decisions made by their peers. This result conveys platform design and policy implications for the next generation of socially-aware investment markets. Such markets will need to explicitly consider the intrinsic value individuals place on cross-referencing their options against the choices made by peers. This social interdependence has the potential to improve welfare, allowing less sophisticated investors to benefit from the wisdom of their peers; however an over-reliance on crowd wisdom can generate conditions under which herding behavior can create systematic inefficiencies.

Similar analyses applied to real-world investment data or field experiments in the spirit of Bursztyn et al. (2014) can add significant incremental value to the literature in this area by contributing evidence of external validity for these findings. A key relationship that cannot be uncovered using the ultimately risk-free CAPS platform is how the contribution of social information towards investor decision making changes in response to different levels of risk exposure. Are investors equally likely to emulate the investment decisions of their peers within a high stakes context? I also believe that similar methodology can be effectively

applied to a number of non-finance contexts where the market actions chosen by individuals are influenced by peer effects and heterogeneous data. As a further application, I present a third paper which applies similar methodology to an entirely different market where the peer groups of video game purchases are shown to strongly influence which titles individuals choose to purchase.

With increasing interconnectivity an increasingly prevalent influence on the daily lives of market actors, research into the understanding how sentiment, socialization, virality of consumption, and peer effects influence the design and outcome of markets will likely emerge as an impactful space for applied econometrics to understand changing patterns of consumer behavior and identify opportunities to remedy inefficient market outcomes.

Chapter 2

DEMAND ESTIMATION FOR SOCIAL VIDEO GAMES: EVIDENCE FROM STEAM

2.1 Introduction

The personal computer (PC) video gaming industry features a rapidly growing market estimated to have generated $36 billion in revenue in 2016.[1] A high-level segmentation of this industry includes specialized hardware, software, and game-related services. Within the software sector there are two major categories of products: games for mobile devices or web browsers and premium PC games. Games in the former category are typically characterized by short development cycles, lower production costs, and a business model which relies on in-app transactions rather than up-front purchase costs. Conversely games in the premium sector, which accounted for an estimated $5.3 billion revenue in 2016, generally feature lengthy development cycles with high costs of production and a traditional pricing model with an up-front purchase cost. This paper focuses on the market for digitally distributed games in this premium sector.[2]

Throughout most of the previous decade, the PC gaming sector featured a conventional physical media supply chain under which development studios partnered with mass-market publishers that specialize in marketing and logistics of physical distribution. In recent years,

[1]Industry evaluation conducted by research firm SuperData based on internal industry tracking data. The evaluation report is available from the following url: https://www.superdataresearch.com/market-data/market-brief-year-in-review/

[2]Henceforth, any discussion of games refers to software within the premium PC sector.

the market has shifted dramatically towards digital distribution which provides customers with a more convenient experience, effectively eliminates the marginal cost of production, and empowers developers to reduce their reliance on publishers. This shift has resulted in a greater variety of games brought to market under a variety of financing models ranging from conventional publisher relationships to independent games crowd-funded through platforms like Kickstarter.[3]

A key driver for this disruption in market structure has been the success of platforms like Steam, a digital distribution platform empowering publishers and developers to sell digital copies of their games, control their own pricing, take advantage of promotional and marketing tools, and deploy software updates to customers. Steam's terms for developers are quite simple; Steam retains 30% of revenue generated by game sale after any value-added tax is applied as a proportional fee in exchange for providing marketplace and support services.[4] In contrast, traditional publishers under physical distribution models typically retained a 70% or larger revenue share. Sellers on Steam retain the ability to choose prices and conduct promotions, but Steam provides advisory services and business intelligence which generate pricing recommendations based on observed demand for existing games with similar features.

2.1.1 Demand Estimation and the Steam Platform

Given the multi-year development cycle for premium games, understanding customer demand is a critical focal point for developers when initially committing resources to a project.

---

[3] In recent years a number of successful Kickstarter projects have succeeded with multi-million dollar funding campaigns serving as a proof of concept for other developers.

[4] There is not an official public policy which confirms this 30% commission; however it is widely confirmed through anecdotal accounts of current and past publishers who have partnered with the Steam platform. Details regarding Steam's public marketplace policy may be found here http://www.steampowered.com/steamworks/FAQ.php

Improvements to this understanding can benefit both producers by reducing risk in the allocation of capital as well as market makers like Steam who seek to drive growth in demand with improvements to platform features. PC games are experience goods, and customer awareness, socialization, and virality matter in determining demand outcomes. This paper seeks to contribute by carefully modeling how dimensions of socialization and peer influence impact demand and interact with pricing strategies. Empirical estimation of demand curves is challenging when observing endogenous equilibrium outcomes. Such exercises rely on observing exogenous supply shifts or the careful use of supply-side instruments in order to identify demand parameters. For many markets, this requirement may be prohibitive of estimating unbiased parameters. The Steam marketplace features several key advantages which provide a structural opportunity to surmount such econometric obstacles.

There are significant benefits to estimation due to Steam's structure as a digital distribution platform which assists with identification of the demand curve. Demand for games on Steam shifts gradually over time as games mature and social perception of quality evolves. Endogeneity emerges as a problem of observing simultaneously determined equilibrium outcomes for prices and quantity demanded. Publishers on the Steam platform are able to price as monopolists facing constant marginal cost of supply; observed price changes represent vertical shifts in a perfectly-elastic supply curve, allowing us to use this price variation to estimate the marginal effect of price changes on quantity demanded after controlling for other determinants of demand. Sellers on Steam frequently conduct sales and promotions, with duration ranging between several hours and several weeks, which generate meaningful price variation and provide identification of price effects conditional on other time-varying

factors.[5] I exploit this platform structure to propose an estimation approach in section 2.3.

2.2 Related Literature

This paper contributes to existing applied micro-econometrics literature with a novel application of individual consumer demand estimation using a rich dataset within an under-studied industry featuring unique demand and cost structures. In particular, this paper focuses on quantifying the impacts of socialization on individual demand which are rarely able to be captured. There are two recent contributions in this area which use network data directly in microeconomic models. Banerjee et al. (2013) who model how awareness and utilization of microfinance loan programs propogate through social networks in rural India. Burstzyn et al. (2014) employ a high-stakes field experiment to model the choices made by managers at a financial brokerage conditional on the strategies employed by their social network.

Challenges exist in claiming causal interpretation of estimated peer effects. Manski (2000) summarizes the social interactions literature and highlights how, even with suitable social data, identification of causal relationships which are attributable to the behavior of peers is difficult due to the "reflection problem" when observed behavior is endogenously determined by membership in one's peer group. A more recent critique of empirical work claiming causal peer effects is provided by Angrist (2014). Advances in the way we describe network structures and manipulate increasingly high quality and fine granularity network data provides some methodological framework for incorporating network and peer effects in frameworks for causal inference. Much of the basis for notation and methodology is detailed in Jackson (2010). Additional approaches to identifying causal effects are proposed in Bramoullé

[5]In addition to publishers exerting autonomy in pricing, the Steam platform itself sponsors several major platform-wide sales during the year which publishers are encouraged to participate in. For example, the "Steam Summer Sale" and "Steam Winter Sale" both involve significant discounts on the price of a large number of titles.

et al. (2009) who use intransitive triads as the basis for identification of network effects. More recently, De Giorgi et al. (2016) provide a theoretical framework for using exogenous shocks experienced by friends-of-friends to provide identification, which they demonstrate using administrative tax records from Denmark.

Previous work infers the existence of network effects on demand using aggregate observational data. Sorensen (2006) identifies a social component to the choice of health plan made by University of California faculty, likewise Patacchini and Venanzoni (2014) idenfity impacts of socialization on the choice of residential neighborhood, lastly Grinblatt et al. (2004) isolate a component of automobile demand as being attributable to the vehicle consumption choices made by neighbors. Additionally, there are set of papers which focus only on estimation of aggregate demand outcomes in cases where those outcomes may be tied to contemporaneous measures of socialization. An example is the work of Enrico Moretti (2008) examining demand for theatrical film releases using box-office data and a social learning model.

I employ an approach to demand estimation motivated by the model specification and notation of Bajari et al. (2015) who provide a survey of modern techniques for demand estimation borrowing both from statistics as well as modern machine learning methods for model selection and estimation. Specific to the market for video games, I find the theoretical structure provided by Bergemann and Vlimki (2006) useful who propose a dynamic model for optimal pricing of experience goods with an extension to include a social learning component in demand. This model differentiates between the optimal pricing strategies employed within mass and niche markets conditional on the degree of social learning through which potential consumers are informed. Similarly, Ishihara and Ching (2012) estimate a dynamic demand model within the used goods market for video games quantifying deterioration of consumption value over time. Zhu and Zhang (2006) estimate a model which quantifies the influence of online customer reviews in the resulting demand for premium video games, partially high-

lighting the important social component to consumption in this market. Overall econometric investigation of the gaming industry is underrepresented in the literature, particularly given the quality of individual-level data which can be gathered from platforms like Steam.[6]

2.3 Modeling Approach

This paper seeks to model customer demand for 11,604 distinct games sold on the Steam platform between January 2014 and March 2017 using a rich dataset of socially-networked customers to understand how individual demand changes in response to the purchase decisions made by friends and peers. I estimate three models; an aggregate demand model which provides a baseline comparison point for elasticities of price and hedonic attributes, an individual-level demand model which excludes the influence of social network measures on the demand decision, and lastly an augmentation of the individual-level model which incorporates customer-specific measures of network depth and activity.

2.3.1 Aggregate Demand

We observe aggregate demand for games $j \in J_t$ where q_{jt} denotes the total number of customers purchasing game j in period t. Individuals may only purchase at most unit of each game, therefore we define ownership as the cumulative summation $Q_{jt} = \sum_{k=0}^{t} q_{jk}$. Additionally, for each title we observe prices p_{jt}, a large number of game-specific characteristics X_{jt} some of which are time varying while others describe hedonic attributes, and measures of community sentiment S_{jt}. I estimate a model for aggregate demand with the following

[6] A characteristic feature of the gaming industry is the heightened emphasis on technological innovation, socialization, and big data which places this market several years ahead of more established retail sectors where the majority of consumption still takes place through physical channels using conventional technologies and data.

general form, borrowing generalized notation from Bajari et al. (2015), where the objective is inference around parameters θ.

$$\ln(q_{jt}) = f(p_{jt}, X_{jt}, S_{jt}, \eta_t, \epsilon_{jt}; \theta) \tag{2.1}$$

The vector η_t captures time-varying seasonality and holiday effects, while ϵ_{jt} represents an idiosyncratic error component. I estimate this model using a flexible linear functional form where:

$$\ln(q_{jt}) = \phi \ln(p_{jt}) + \beta_X' X_{jt} + \beta_S' S_{jt} + \eta_t + \lambda' \tau \left(X_{jt}, S_{jt}, \eta_t\right) + \epsilon_{jt} \tag{2.2}$$

In this specification τ represents an interaction operator which generates many-to-many interactions between prices, game characteristics, and sentiment measures. The parameter ϕ provides a baseline estimate of price elasticity which is further modified by components of the parameter vector λ which apply to interactions of price with other observables. Results from this model specification are presented in Section 2.5.1.

2.3.2 Individual Demand

At the individual level, we observe customers $i \in N_t$, where N_t grows as new users register on the Steam platform. For each customer we observe discrete demand $q_{ijt} \in \{0, 1\}$ which represents a purchase of game j made by customer i in period t. Similarly, we can express individual ownership of a game j as the cumulative summation $Q_{ijt} = \sum_{k=0}^{t} q_{ijk}$. To model the demand decision at the individual level I employ a standard latent utility framework for discrete demand, where individual i's unobserved utility u_{ijt} from consuming a unit of game j is a function of price p_{jt}, characteristics of the game X_{jt}, individual-specific heterogeneity D_{it}, measures of network socialization and peer influence S_{ijt}, and time period or seasonality

effects η_t.

$$u_{ijt} = f(p_{jt}, X_{jt}, D_{it}, S_{ijt}, \eta_t, e_{ijt}; \theta) \tag{2.3}$$

I approximate this unobserved utility using a linear model specification similar to the aggregate specification in equation 2.2 and estimate the parameters θ by maximum likelihood using a logit model where:

$$\begin{aligned} \tilde{u}_{ijt} &= \phi \ln(p_{jt}) + \beta_X' X_{jt} + \beta_D' D_{it} + \beta_S' S_{ijt} + \eta_t + \gamma_j + \lambda' \tau \left(X_{jt}, S_{ijt}, \eta_t\right) \\ \Pr(q_{ijt} = 1) &= \frac{\exp(\tilde{u}_{ijt})}{1 + \exp(\tilde{u}_{ijt})} \end{aligned} \tag{2.4}$$

Results from this model specification are presented in section 2.5.2. I present results which first exclude S_{ijt} and ultimately add these covariates to augment the baseline model with measures of network socialization and peer influence.

2.3.3 Social Network Data and Demand

The observed Steam network is represented as a graph on a set of N nodes, each of which represents a distinct platform user i. The network at time period t is defined as an undirected dynamic graph $(N_t, \boldsymbol{g}_t)$ where $\boldsymbol{g}_t$ is the time-varying adjacency matrix on the set of nodes where each element g_{ijt} indicates the relationship between users i and k in period t. We follow Jackson (2010) and denote $g_{ikt} = 1$ as $ik \in \boldsymbol{g}_t$. In order to render the information contained in the network tractable for modeling, we follow the standard practice to reduce dimensionality to a number of key descriptive statistics.

Each period, we can describe the set of neighbors for node i as $N_{it}(\boldsymbol{g}_t) = \{k | ik \in \boldsymbol{g}_t\}$. We measure the degree of each node i as its number of neighbors in the network, such that $d_{it} = |N_{it}(\boldsymbol{g}_t)|$. For each link ij we measure support as the set of links $S_{it}(\boldsymbol{g}_t) = \{ik | kl \in$

$\boldsymbol{g}_t$ and $il \in \boldsymbol{g}_t$ where $k \neq l\}$ where nodes i and k share a mutual connection with node k. We can therefore express the fraction of node i's connections which are supported as $s_{it} = |S_{it}(\boldsymbol{g}_t)|/d_{it}$. Lastly, we measure the set of second degree connections which represent friends-of-friends in the Steam context. We define $N^2_{it}(\boldsymbol{g}_t) = \{\cup_{k \in N_{it}} N_{kt}\} - N_{it}$ where the second degree $d^2_{it} = |N^2_{it}(\boldsymbol{g}_t)|$ counts the number of nodes which are connected to a neighbor k of node i but not to node i directly.

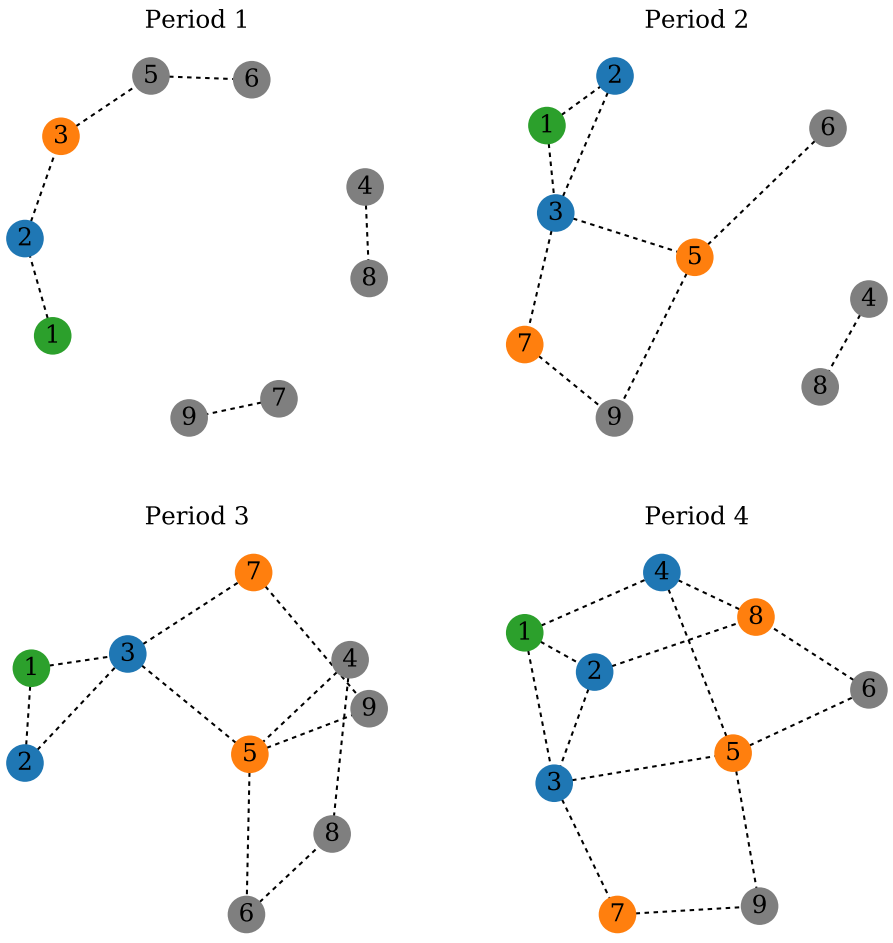

Figure 2.1: Example Network Structure

As an illustrative example, Figure 2.1 depicts the dynamic transition of a simple graph $\boldsymbol{g}_t$

over four periods. We anchor the example to node $i = 1$, depicted in green. The node's neighbors $N_t(\boldsymbol{g}_t)$ are drawn in blue, second-degree neighbors $N_t^2(\boldsymbol{g}_t)$ in orange, with higher-order neighbors drawn in grey. In the initial period, there are only 5 nodes in the network, while additional nodes i enter in subsequent periods. For node $i = 1$, the network grows in degree over time from $d_{1,1} = 1$ to $d_{1,4} = 3$. Table 2.1 reports these statistics for each example network state. We isolate changes in the degree, support, and second degree of each user's network to test hypotheses about how the richness of one's network and the information which flows through it influences the demand decision in the Steam market for video games. The network structures in our actual data are substantially more complex and described in section 2.4.1.

Table 2.1: Example Network Statistics - Node 1

Period t	Degree	Degree^2	Support
1	1	1	0.00
2	2	2	1.00
3	2	3	1.00
4	3	3	0.66

We can augment this structure to consider the interaction of these descriptive network measures with the demand decisions made by an individual's neighbors $N_{it}(\boldsymbol{g}_t)$, their friends-of-friends $N_{it}^2(\boldsymbol{g}_t)$, and their level of network support $S_{it}(\boldsymbol{g}_t)$. For both purchases and ownership, the observed demand outcomes define vectors $\boldsymbol{q}_{jt}$ and $\boldsymbol{Q}_{jt}$ with length N_t. We compute the interaction of these demand vectors with the adjacency matrix $\boldsymbol{g}_t$ to compute the network statistics of degree, second degree, and support for each node i within the set of neighbors N_{it} who are also purchasers or owners of game j. Specifically we can compute the number of purchasing neighbors as $dq_{ijt} = |\{k|ik \in \boldsymbol{g}_t \text{ and } q_{kjt} = 1\}|$ and the number of owning neighbors as $dQ_{ijt} = |\{k|ik \in \boldsymbol{g}_t \text{ and } Q_{kjt} = 1\}|$. These statistics can be efficiently com-

puted by multiplying the demand vectors $\boldsymbol{q}_{jt}$ and $\boldsymbol{Q}_{jt}$ by the adjacency matrix $\boldsymbol{g}_t$ and the support matrix $\boldsymbol{s}_t$. The addition of these measures of socialization to the demand model is the central contribution of this paper, results of which are presented in section 2.5.2.

2.3.4 Primary Hypotheses

Using this modeling framework we can test a number of compelling hypotheses. A natural starting point is to quantify the price elasticity of demand for each observed price band. We would certainly expect to see negative price elasticities, however we consider the hypothesis of whether price is equally elastic for games released at the premium end of the pricing range relative to cheaper titles. Regression results from the aggregate demand specification are reported in Table 2.4 on page 57. In addition to the effect of price, these results allow for an examination of how common product characteristics influence product demand within each distinct pricing segment.

Furthermore, we closely examine hypotheses around socialization and demand, both at the aggregate and individual levels. We evaluate whether aggregate community sentiment in the forum of user-submitted reviews, YouTube videos, and Twitch broadcasts manifests a meaningful causal impact on observed demand. At the individual level, we test whether measures of network richness and saturation manifest a material increase in a customer's likelihood of purchasing a title. Individual level demand estimates are reported in table 2.5 on 64. We test this paper's primary hypothesis that peer consumption bears powerful influence over individual demand decisions against the null hypothesis that measures of social network engagement have no effect on the individual purchase decision.

2.4 Steam Platform Data

Our data is extracted from three complementary sources which provide both aggregate and individual-level views into the demand for games within the Steam platform. Our primary sources for Steam platform data are the official Steam Web API and Steam Store API.[7] These API endpoints provide player-level data regarding the library and social connections of each player as well as game-specific data which characterizes a title and its features. Using these data sources I have collected data for 91,693 active Steam users whose game libraries span 11,604 distinct game titles.

This data forms an unbalanced panel where users i enter the dataset upon the date of their registration with the Steam platform and games j enter the data upon the date of their initial release. I assume no attrition in i or j, which appears to be a reasonable assumption given the limited scope of evaluation in this paper, however if conducting analysis over a longer time period it would be prudent to cull no-longer-active users and games from the data.

2.4.1 Individual Demand Data

For each of the 91,693 sampled Steam users, we observe the purchase and game-play history for the titles in that player's library. Figure 2.2 depicts the purchase dates, titles, and prices of the game library for an example user. We observe a broad range of prices at which titles

[7] Details regarding the Steam API are available here: https://steamcommunity.com/dev. Use of the API requires registering an account for an API key which can be used to access a variety of data endpoints which return JSON objects. Game-specific characteristic features are obtained from the Steam Store API which differs slightly from the Steam Web API. For this project, I used Python to automate the retrieval of data from these official API endpoints.

are sold.[8] Examining this price distribution suggests clear clustering into several price bands ranging from *premium* for games which retail for $45 or more, *mid-tier* games which feature prices between $20 and $35, and *budget* titles which sell for $15 or less. The Steam API does not report the price paid by the user, but we observe the full history of daily mean prices for each game which allows us to infer the price paid using the purchase date of the title. The mean user owns 10.2 games purchased at an average price of $31.11.

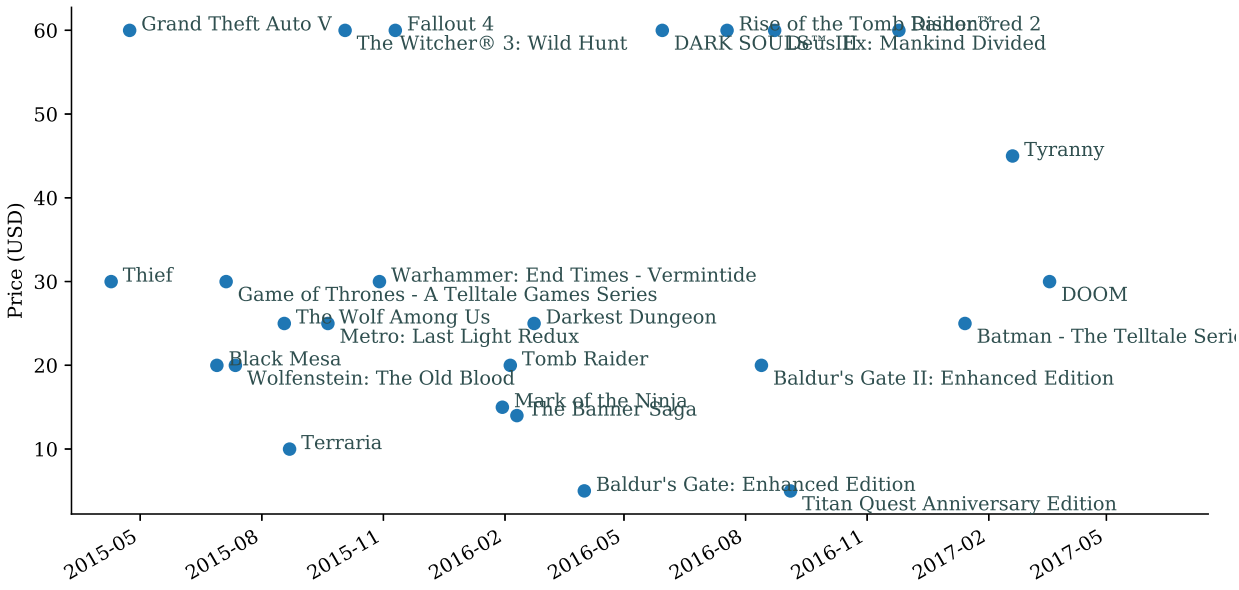

Figure 2.2: Steam Individual Demand - Titles and Prices

The most noteworthy data element we observe at the individual level is the time-varying

[8] Observing the exact date of a purchase using the available Steam Web API is unfortunately somewhat imprecise. The ideal method for inference regarding the purchase date is to observe the user's library for sequential days and infer purchases on day t as the set-wise difference from the library in period t and period $t-1$. I employ this strategy where possible, however each user's library was typically snapshot on a weekly basis to reduce the volume of API calls required to generate the dataset. Where such a gap exists I attempt to refine identification of the purchase date using the time when the user first plays a game. This is not always accurate, as many gamers may purchase games when they are on sale, but delay playing them until a later date. The overall quality of the observed data is quite high, but some measurement error in the exact purchase date and therefore the price paid does exist.

social network for each user. On the Steam platform social connections take the form of friendships which are bilateral connections agreed upon by both parties.[9] For each user i, we observe the peers with whom that user has a friendship as well as the date on which that friendship was created. This allows us to precisely define the edges $\{g_{ijt}\}$ and neighboring nodes N_{it} of each user's network at each point in time. This data provides a platform upon which to evaluate peer influence and network effects with respect to individual demand outcomes. Figure 2.3 depicts the social network at the end of 2016 for the same example user, depicted in green at the center of the network graph.

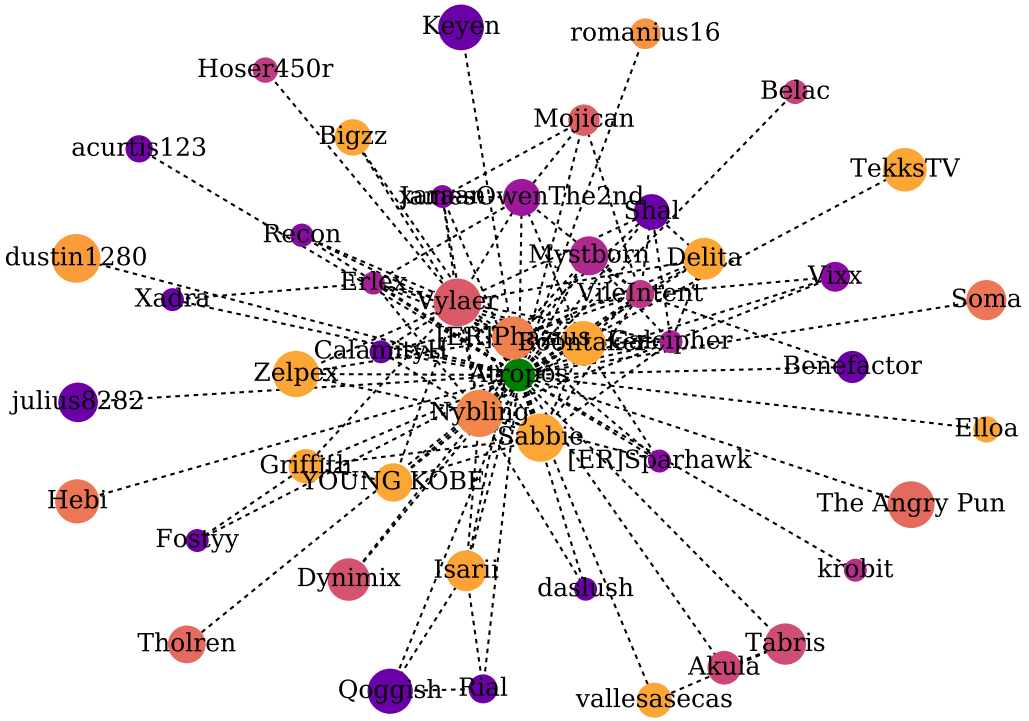

Figure 2.3: Steam Individual Social Network

[9] This type of social connection differs from "favorites" or "followers" used commonly on social media sites like Twitter, YouTube, or Instagram. Friendships on Steam require one user to initiate the request and the other to approve it before a connection is established.

In this visualization, we show only the first-degree neighbors N_{it} of the target node i. The radius of each neighboring node k is drawn proportionally to the number of titles in that player's game library Q_{kjt}. The color intensity of each node represents their number of second-degree neighbors N_{kt}, indicating each node's level of socialization where warmer colors (orange and red) depict highly socialized players while cooler colors (purple and blue) denote users with fewer friendships themselves. This particular user has 47 friendship connections, $d_{it} = 47$. Of these 47 connections, 24 are supported by a mutual friendship, $s_{it} = 0.511$. Furthermore, these 47 friends feature a collective second-order network of 820 friends-of-friends, $d^2_{it} = 820$. We compute this summary data for each user $i \in N$ for each period $t \in T$. For clarity of interpretation, cross-sectional summary statistics from March 2017 are presented in the Table 2.2.

Table 2.2: Steam Network Descriptive Statistics

	Degree - N_{it}	Degree2 - N^2_{it}	Support - s_{it}
Mean	137.65	5,868.06	0.647
Std. Dev.	108.89	5719.65	0.338
Median	127	4591.00	0.786
Min	0	0	0.000
Max	1,035	42,357	1.000
Observations	91,693		

Statistics are reported for the end-point of the observed sample in March 2017.

2.4.2 Game Characteristics and Features

To support this demand data we collect a large number of covariates X_{jt} which describe game j and contextualize its demand. Core descriptors which may affect demand include the developer, publisher, genre, and ESRB rating of the title. Certain genres are generally more popular, specific developers have strong reputations for quality, and ESRB restrictions

may restrict (or enhance) a game's target audience. In addition to these core descriptors we observe a large amount of unstructured data for each game including a variable number of category associations, user-specified text tags, and a rich text description of the title containing text and images. We parse these unstructured fields using natural language processing methods to reduce its dimensionality and extract additional characteristics. Across all titles we observe 5,036 publishers, 6,681 developers, 30 genres, 32 categories, and 1,766 user submitted tags. Table 2.3 presents example data for *Doom*, a popular science-fiction shooter which first released in 2016.

Table 2.3: Example Game Characteristics - *Doom* (2016)

Characteristic	Values
Initial Price	59.99 USD
Release Date	2016-05-12
Publisher	Bethesda Softworks
Developer	id Software
Genres	Action
Features	Single-player, Multi-player, Co-op, Controller Support, Achievements, Trading Cards, Steam Cloud
Platforms	Windows
Languages	English, French, Italian, German, Spanish, Japanese, Polish, Portuguese-Brazil, Russian, Traditional Chinese
ESRB Rating	M (17+)
User Tags	FPS (1204), Action (1080), Gore (1058), Demons (887), Shooter (837), First-Person (769), Multiplayer (726), Great Soundtrack (648), Single-player (628), Fast-Paced (609), Sci-fi (588), Classic (548), Horror (547), Atmospheric (488), Difficult (359), Remake (289), Zombies (242), Co-op (144), Blood (99), Memes (68)

The integer following each User Tag denotes the number of Steam users who associated this tag with the game when submitting a review.

Game features designate whether the game includes single-player, multi-player, or co-op

support (or some combination of the three) and imply whether consumption of the game is an explicitly social experience. Achievements and Trading Cards represent specific features of the Steam platform with which game developers can support integration by defining in-game achievements for players to attain as well as collectible trading cards which can be exchanged with other players on the Steam Marketplace. Games which offer these features provide increased consumption value for players on the Steam platform in addition to the pure game characteristics.

2.4.3 Aggregate Demand and Community Sentiment

The third data component which empowers our aggregate-level analysis is data acquired from the community-managed service SteamSpy which conducts daily crawls of the Steam database. The SteamSpy site collects daily price data as well as precise estimates of aggregate demand using a representative random sample of individual Steam users.[10] From SteamSpy, we collect aggregate demand, incremental purchases, daily prices, and measures of player activity. SteamSpy also provides other aggregate measures of community engagement including the number of game-related videos and views on YouTube, the number of live channels and viewers on Twitch, and the number of new user reviews published on the Steam store page.[11]

Figure 2.4 on the following page depicts example aggregate demand data for *Doom*. We observe cumulative ownership Q_{jt} growing over time with volatility clustering of purchases

[10] Additional details regarding SteamSpy are available on the "About" page http://steamspy.com/about. SteamSpy uses a random sampling strategy to infer the total number of Steam users who own and play each title on Steam. A description of the sampling methodology and statistical power is provided in Kyle Orland's 2014 *Ars Technica* article.

[11] Twitch (https://twitch.tv) is a popular live-streaming platform where gamers can broadcast live gameplay and interact with their viewer communities.

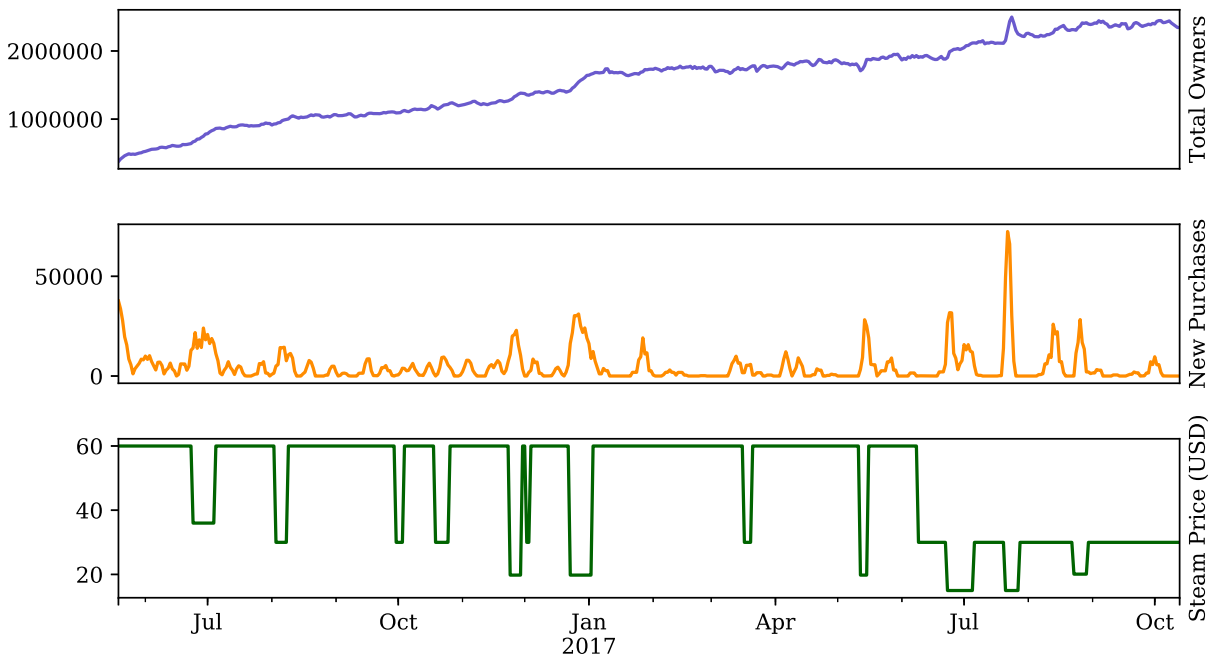

Figure 2.4: Daily Aggregate Demand and Sales Price - *Doom* (2016)

q_{jt} where heightened quantity demanded visually correlates with periodic pricing promotions. For most games on Steam, we observe stable prices interspersed with periods of significant (typically 10% or greater) discounting. These sudden and meaningful variations in price allow us to identify price elasticity of demand since they are not motivated by changes in the supply curve. We summarize the distribution of initial retail prices and observed discount rates in Figure 2.5 which illustrates banding in initial price (MSRP) as well as significant variation in discounting relative to this initial price.

The final data element yielded by the SteamSpy API contains observed measures of community sentiment. These characteristics are depicted in Figure 2.6 for the example case of *Doom*. We observe on a daily basis the number of users submitting positive and negative reviews of each game. Reviews may only be submitted by individuals who have purchased

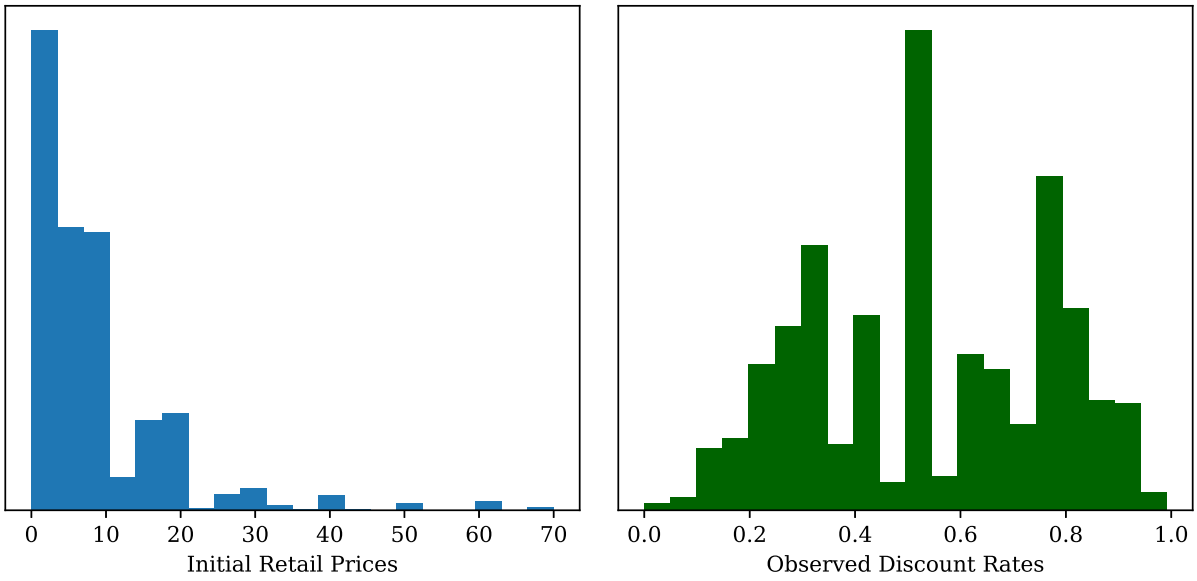

Figure 2.5: Initial Prices and Observed Discount Rates

the game and the reviewer must specify whether the review is positive or negative, removing the need for subjective interpretation of the provided text. We also observe the number of Twitch broadcasters actively streaming for this game, and the number of new views for YouTube videos tagged to the game. We include these measures of community sentiment as additional time-varying controls in our models and evaluate the hypotheses that heightened levels of community sentiment generate increases in demand for each game. These aggregate measures of community sentiment complement the individual-specific measures of peer influence described in Section 2.4.1.

2.5 Econometric Results

The following sections present results of the model specifications outlined in Section 2.3; an aggregate demand specification and an individual-level discrete-choice logistic specification

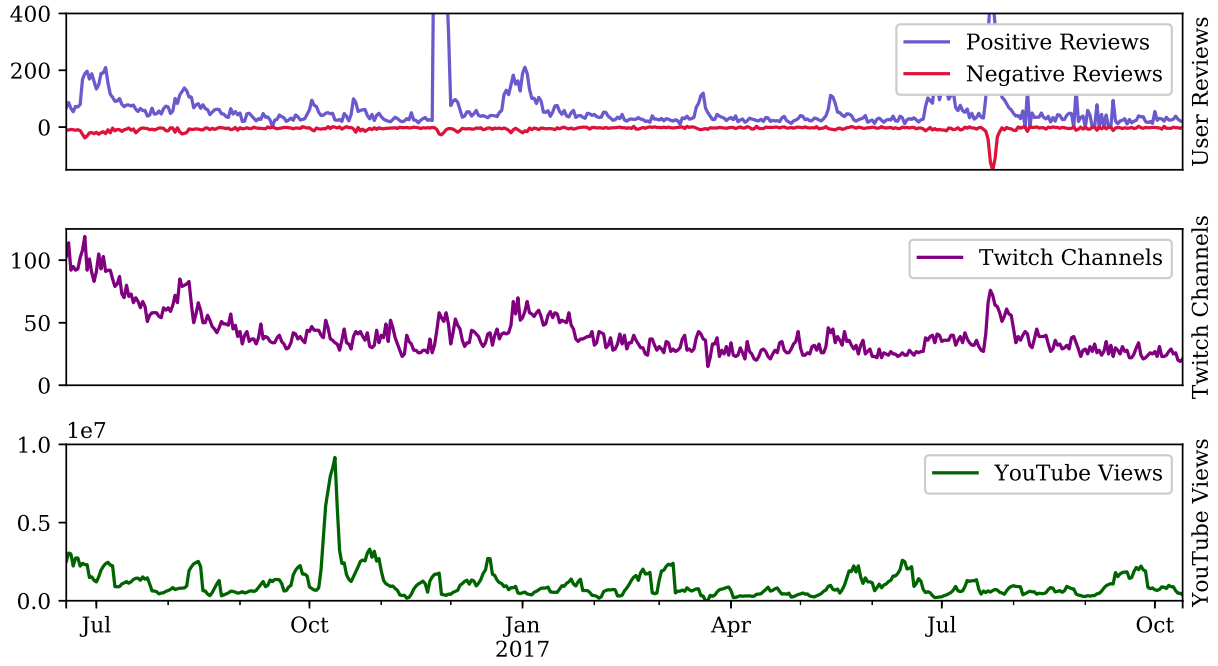

Figure 2.6: Measures of Aggregate Sentiment - *Doom* (2016)

augmented with measures of socialization and peer influence. In all models I control for variation in price, game attributes, community sentiment, seasonality, and interaction terms of attributes. Each sample is stratified into price bands according to key breakpoints in the distribution of initial recommended retail price. I estimate a separate model for each price band to better characterizes the price elasticity along different segments of the demand curve.

2.5.1 Aggregate Demand Estimates

For the aggregate demand case, I estimate the linear regression model specified in Section 2.3.1 in which I regress log-demand $\ln(Q_{jt})$ on log-price $\ln(p_{jt})$ and a large variety of controls including game characteristics X_{jt}, measures of aggregate sentiment S_{jt}, seasonality, and

interactions of these features. In total, I consider a design matrix including 24,604 potential covariates which are considered for the model. To reduce dimensionality for estimation and restrict the model to the most meaningful explanatory factors, I employ a first stage LASSO regression in which I train the model on a $3/4^{\text{ths}}$ sample of the 7,045 games and reserve the remaining games for cross-validation. Defining the design matrix of all possible features as $\boldsymbol{Z_{jt}} = \{\ln(p_{jt}), X_{jt}, S_{jt}, \eta_t, \tau(\ln(p_{jt}), X_{jt}, S_{jt})\}$, I use cross validation to solve for the optimal LASSO parameter λ where:

$$\min_{\theta} \quad \left\{ \frac{1}{N} ||q_{jt} - \boldsymbol{Z_{jt}}\theta||_2^2 + \lambda|\theta|_1 \right\} \tag{2.5}$$

The second stage model includes the set of features from $\boldsymbol{Z_{jt}}$ where the first-stage parameters θ are non-zero. The primary OLS regression on this subset of selected features is reported in Table 2.4 on pages 57 and 58.[12] Variable definitions for included features are available in Appendix .4. Summary statistics are reported in Appendix .5.

We estimate price elasticities of demand which range from a relatively elastic -1.3497 in the upper-most price band ($44.99+) to a very inelastic -0.1621 in the lowest ($0.99 to $4.98) price band. This deterioration of price elasticity is expected, but it particularly noteworthy that we only observe evidence of elastic demand for the most expensive games on the Steam platform. This suggests that despite having copyright over a game's intellectual property, sellers on Steam do not always appear to price as monopolists, increasing price when demand is inelastic. The estimated elasticity for premium games is somewhat surprising in that these titles have arguably fewer good substitutes. For example, a game like *Doom* (initially priced at $59.99) which belongs to a celebrated franchise and is created by a well-respected developer has a very limited cohort of similar games which could realistically be considered substitutes.

[12]Features required for the testing of primary hypotheses are always included in the second-stage model.

Table 2.4: Aggregate Demand Estimates

	Price Band				
Covariate on $\ln(q_{jt})$	\$44.99	\$29.99	\$14.99	\$4.99	\$0.99
Log Price	-1.3497*	-0.6472*	-0.4354*	-0.1715*	-0.1621*
	(0.1374)	(0.1364)	(0.0373)	(0.0099)	(0.0087)
Ownership Share	-0.0063	-0.0125	0.0042*	0.0118*	0.0261*
	(0.0053)	(0.0089)	(0.0016)	(0.0027)	(0.0055)
Game Age (Months)	-0.0248*	-0.0145*	-0.0020*	-0.0025*	-0.0013
	(0.0053)	(0.0032)	(0.0010)	(0.0005)	(0.0007)
Median Hours Played	-0.0015*	-0.0018	-0.0007	-0.0014*	-0.0011*
	(0.0007)	(0.0012)	(0.0005)	(0.0003)	(0.0004)
Supported Languages	-0.0239	-0.0236	0.0179*	0.0056*	-0.0003
	(0.0393)	(0.0526)	(0.0038)	(0.0015)	(0.0019)
Recent Positive Reviews	0.0003*	0.0007*	0.0010*	0.0017*	0.0080*
	(0.0001)	(0.0003)	(0.0004)	(0.0004)	(0.0011)
Recent Negative Reviews	-0.0063*	-0.0168*	-0.0073	-0.0106	-0.0257
	(0.0018)	(0.0068)	(0.0075)	(0.0064)	(0.0153)
Recent Sentiment	-0.0007	-0.0004	0.0025*	0.0023*	0.0023*
	(0.0004)	(0.0004)	(0.0002)	(0.0001)	(0.0002)
Cumulative Sentiment	0.0015	0.0002	-0.0000	0.0001	-0.0000
	(0.0012)	(0.0004)	(0.0000)	(0.0000)	(0.0000)
Twitch Viewers	0.0208*	0.0150	0.0341*	0.0344*	0.0210
	(0.0074)	(0.0105)	(0.0102)	(0.0065)	(0.0158)
Twitch Channels	0.0002	0.0182*	0.0009	0.0216*	-0.0276*
	(0.0003)	(0.0088)	(0.0010)	(0.0027)	(0.0084)
YouTube Views	0.0097*	0.0125*	0.0163*	0.0076*	0.0018*
	(0.0020)	(0.0030)	(0.0045)	(0.0010)	(0.0003)
YouTube Videos	0.0007	0.0064*	0.0139*	0.0018*	0.0011
	(0.0016)	(0.0014)	(0.0014)	(0.0007)	(0.0010)
Games	163	167	1,007	3,290	2,418
Observations	47,359	63,234	410,292	1,427,445	1,211,564

Table continued on the following page.

Prices in each column denote the right-endpoint of each price band.

Standard errors are clustered by title.

Coefficients marked (*) are statistically significant at the 5% level.

Table 2.4: Aggregate Demand Estimates - Continued

	Price Band				
Covariate on $\ln(q_{jt})$	\$44.99	\$29.99	\$14.99	\$4.99	\$0.99
Single-Player	0.1896*	0.2095*	0.0326	-0.0024	0.0573
	(0.0146)	(0.0434)	(0.0885)	(0.0307)	(0.0545)
Multi-Player	-0.2184*	0.0494	0.1019*	0.0509*	-0.0075
	(0.0295)	(0.0304)	(0.0421)	(0.0200)	(0.0315)
Cooperative	0.2760*	-0.2920*	0.0319	0.0338	-0.0499*
	(0.0340)	(0.0100)	(0.0390)	(0.0177)	(0.0245)
Massively Multi-Player	-0.5199*	-0.4142*	-0.0338	-0.2726*	-0.0765
	(0.0659)	(0.0976)	(0.2436)	(0.0973)	(0.1286)
Virtual Reality	-0.1647*	0.1909*	0.1963*	0.0370	0.0036
	(0.0133)	(0.0413)	(0.0770)	(0.0412)	(0.0663)
Trading Cards	-0.6535*	0.1475*	0.0171	-0.0098	0.0998*
	(0.0901)	(0.0205)	(0.0321)	(0.0117)	(0.0125)
Tracked Stats	0.1665*	0.1855*	-0.0356	0.0084	-0.0024
	(0.0240)	(0.0505)	(0.0231)	(0.0085)	(0.0093)
Mod Support	-0.9318*	-0.1589	0.1235*	0.0841*	-0.0328
	(0.1219)	(0.3196)	(0.0454)	(0.0214)	(0.0327)
Has Achievements	-0.1963*	0.0908*	0.0900*	-0.0320	-0.0404*
	(0.0268)	(0.0257)	(0.0442)	(0.0170)	(0.0175)
LASSO Features (Included)	Most Significant Features				
Day-of-Week (5 of 7)	Friday$^+$, Saturday$^+$, Sunday$^+$, Thursday$^-$				
Month (10 of 12)	December$^+$, June$^+$, April$^+$, August$^-$				
Genre (24 of 32)	RPG$^+$, Sexual Content$^+$, Sports$^+$, Action$^-$, Indie$^-$				
Tag (92 of 337)	Fast-Paced$^+$, Rogue-like$^+$, Arcade$^+$, Violent$^+$, 4-Player Local$^-$, Mystery$^-$, Sci-Fi$^-$, Rogue-lite$^-$, Dinosaurs$^-$				
ESRB Rating (4 of 6)	Mature$^+$, Adult$^+$, Everyone$^+$				
Publisher (41 of 5,036)	Paradox Interactive$^+$, Square Enix$^+$, THQ$^+$				
Developer (27 of 6,681)	Bethesda Softworks$^+$, Ubisoft Montreal$^+$, Rockstar Games$^+$				
Interactions (201 of 10,942)	(Recent Review Sentiment$\times$June)$^+$, (RPG$\times$Multi-Player)$^+$, (Twitch Viewers$\times$Shooter)$^+$				
Games	163	167	1,007	3,290	2,418
Observations	47,359	63,234	410,292	1,427,445	1,211,564

Prices in each column denote the right-endpoint of each price band.
Standard errors are clustered by title.
Coefficients marked (*) are statistically significant at the 5% level.

However, the elasticity estimated for these premium titles suggests that customers, in general, are more price sensitive within this premium segment whereas within lower-priced segments, customers are measurably less motivated by discounts.

In addition to the first-order effect of price on quantity demanded, these results identify a number of key factors which shift the demand curve. We observe an insignificant or positive impact of ownership share, defined as the number of total owners divided by the number of active players. This estimate is subject to competing underlying forces. Since each individual may at most consume a single unit of a particular game, as more players have already purchased a title the eligible demand pool contracts. Counter-acting this diminishing of the eligible customer pool is the effect of virality within the player community; as a greater proportion of active players purchase and play a certain title, the halo effect due to word-of-mouth and peer influence may cause increases in demand. Indeed, we observe that for lower price bins, the net effect of expanding ownership is positive.

This is structurally consistent with the nature of consumption goods where each individual consumes at most a single unit. As more players have already purchased a good the eligible demand pool of active players shrinks. Additionally, we observe a clear negative relationship between the age of a game and it's demand. Newer games tend to be more popular, have improved graphical fidelity, and support more modern game-play features. Interesting, the marginal value of high-priced games appears to depreciate more rapidly with age than for lower-priced titles where demand decreases by roughly 2.5% for each month after a game's initial release. We estimate a consistently negative impact of median hours played, which describes the typical amount of time existing owners of a game have spent playing it, which suggests that consumers tend to prefer shorter games. This represents a somewhat counter-intuitive finding and merits further investigation when contrasting these results with those from the individual-level model.

The included measures of community sentiment reveal interesting, but not entirely surprising, impacts of social perception on realized demand. Each additional positive game review submitted by a game owner during the previous 7 days increases resulting demand, while we observe decreases in demand due to recent negative reviews. Furthermore, the overall proportion of positive-to-negative reviews submitted within the prior week tends to increase demand. While this is unsurprising as reviews are correlated with unobserved product quality, the relationship between reviews and price bands is very interesting. We see clearly that user-submitted reviews are considerably more influential for games in lower price bands while exhibiting relatively little impact on the demand for premium games.[13] In contrast to the measurable influence of recent review sentiment, a game's cumulative proportion of total reviews bears almost no impact on realized demand. This reveals that gamers, in practice, care very little about the historical quality of a game but place far more emphasis on it's current reputation. It is not uncommon for games to undergo significant upgrades or transformations after being released as developers add new content, fix reported issues, and improve game performance.[14]

Additional community sentiment measures from Twitch and YouTube suggest, unsurprisingly, that heightened levels of exposure on those popular media platforms are beneficial for a game's demand. In general, we observe that the number of people engaging with those platforms (viewers and views) tends to be more impactful than increases in the amount of

[13]This draws similarity to the anecdotal phenomenon that moviegoers are more likely to be influenced by critical reviews for an independent film than for a big summer blockbuster.

[14]In fact, the ability to easily deploy such upgrades in the form of "patches" to game owners is one of the hallmark advantages of digital distribution platforms like Steam. Under historical physical distribution models, developers had limited ability to improve the quality of their software post-release. Patching was only possible as an opt-in mechanism for the technologically sophisticated gamer. Today, game improvements are pushed and applied automatically through platforms like Steam. A growing concern within the gaming community is that this ease of deployment has created a moral hazard problem for publishers and developers who are now incentivized to release games prematurely.

game-specific content which is available (channels and videos).

Additionally, our model includes a large number of features reported in the second half of Table 2.4 on page 58. Of the primary game features included in the model we observe that single-player games are more highly demanded within higher price bands, but multi-player games have increased popularity within lower price ranges. This suggests that if a candidate game benefits from the participation of one's friends, lower price points may present a more attainable threshold towards capturing that game's target experience. Massively multi-player games, which generally require considerable time investment by their players are less demanded in all price bins. Virtual reality support is an emerging feature which has not yet effectively penetrated the premium-priced segment, but is estimated to contribute an approximate 20% increase in demand for moderate price bands. Aspirational features like trading cards, tracked statistics, and achievements are each estimated to have mixed effects by price band without evidence for a unifying hypothesis of their influence.[15]

Lastly, we characterize the influence of features identified by the first-stage LASSO selection. The full set of estimated features is too cumbersome to report directly, so I report several of the most significant features within each category to provide intuition for the richness of controls included in the model. The model identifies significant weekly seasonality effects for weekend demand, and strong annual seasonality in June and December.[16] We identify the 24 most (or least) popular genres like "Role-Playing Game" and "Sports" as well as 92

[15]These game features are ancillary to the actual game-play itself and describe Steam platform meta features which allow gamers to track and share their performance with the broader Steam community. Trading cards are used to redeem "badges" which can be displayed on one's user profile. Achievements represent a tracked list of in-game challenges which can be completed and displayed. Games which support tracked stats will monitor and share player performance data, allowing them to compete on leader-boards for community prestige.

[16]These months capture the timing of the Steam Summer Sale and Winter Sale respectively, which are major recurring events which feature considerable price promotions accompanied by heightened advertising and community events.

prominent user-submitted tags such as "Rogue-like" and "Violent".[17] We observe that games which have an ESRB rating of Mature or Adult are relatively more popular, as are games rated as suitable for children. Furthermore, the LASSO regression identifies 27 developers, 41 publishers, and 201 interaction terms which are estimated to generate meaningful positive or negative impacts on resulting demand.

2.5.2 Individual-Level Demand Estimates

I estimate individual customer demand as described in Section 2.3.2 by generating a panel of users i, games j, and weeks t. For the individual level-estimates we examine weekly demand to assist with the feasibility of estimation. I employed a sampling logic in collecting individual level data in which we randomly sampled 1,461 users who satisfied minimum activity and profile visibility requirements.[18] These 1,461 seed users featured a combined set of 91,693 neighbors who in turn have a broad network of friends-of-friends which, in total, defines an observed network $\boldsymbol{g_t}$ containing 10,786,314 nodes.

We observe demand data for the set of seeds i and their first-degree neighbors N_{it}. We construct a dynamically balanced panel for of seed users i where for each seed we observe the users network degree d_{it}, second degree d_{it}^2, and support s_{it}. Furthermore, we observe the extent to which a game j is demanded within each users set of neighbors. This generates game-specific measures of network richness including peer demand dq_{ijt}, peer ownership

[17] Amusingly, the user-submitted tag "rogue-lite" was identified as a strong negative effect. This is a derisive term for a game which attempts to feature "rogue-like" game-play but executes poorly upon that design objective.

[18] Specifically, I used a three-stage sampling strategy. In the first stage I randomly sample integer Steam IDs without replacement and identify 100,000 existing users. Using this first-stage sample, I restrict to users who have non-private user profiles and some measurable market activity more recently than January 01, 2015. This generates an eligible set of seed users, of whom there are 87,324. From this set of eligible seeds, I iteratively drew, without replacement, individual users who became "seeds".

dQ_{ijt}, supported demand sq_{ijt} and supported ownership sQ_{ijt}. The most significant technical achievement presented in this paper involves the careful construction of the design matrix which includes these network features which collectively are added to the logistic demand regression as the set of social covariates S_{ijt}. Crucially, these social features are computed at the daily level and lagged prior to weekly aggregation, so covariates like supported demand reflect peer purchases which occurred strictly prior to the individual demand decision being modeled. The final dataset used in estimation includes 104,558,469 observations of individual demand outcomes.

For feasibility in estimation, we rely upon the same set of selected features and interactions chosen by the aggregate first-stage LASSO regression from Section 2.5.1 and apply this same feature set in our individual-level model. The dimensionality reduction from the first-stage linear model is critical given the size of the individual-level dataset in order to feasibly estimate the logistic parameters by maximum likelihood. I estimate the logit model described in equation 2.4 including the observed weekly demand of 1,461 seed users between January, 2016 through April, 2017. Results are presented in three parts in Table 2.5 on pages 64 through 66.

These individual-level regression estimates complement the aggregate results presented previously to clarify what appear to be several powerful demand relationships within the data. Firstly, the price elasticity estimates are comparable to those from Table 2.4, however less elastic in the upper-most price band where, at the individual level, I estimate a price elasticity of -0.7576 in contrast to the more elastic -1.3497 from the aggregate results. This suggests that some responsiveness of total demand is likely attributable to microeconomic factors which are correlated with price changes but unobserved in the aggregate model. I will return to this hypothesis when discussing the estimated social network effects. In keeping with the aggregate results, I estimate progressively more inelastic demand for lower price

Table 2.5: Individual Logit Demand Estimates

	Price Band				
Covariate on $\Pr(q_{ijt} = 1)$	\$44.99	\$29.99	\$14.99	\$4.99	\$0.99
Log Price	-0.7576*	-0.6670*	-0.6064*	-0.1130	-0.0645
	(0.0904)	(0.1291)	(0.0773)	(0.1160)	(0.1527)
Ownership Share	0.0273*	0.0117	0.0221*	0.0506*	0.0785*
	(0.0030)	(0.0075)	(0.0031)	(0.0047)	(0.0187)
Game Age	-0.1132*	-0.0603*	-0.0397*	-0.0411*	-0.0250
	(0.0042)	(0.0052)	(0.0029)	(0.0053)	(0.0131)
Median Hours Played	-0.0307*	-0.0780*	-0.0145	-0.1470*	-0.0180
	(0.0020)	(0.0081)	(0.0080)	(0.0304)	(0.0421)
Supported Languages	0.1292*	0.0335*	0.0112	-0.0023	-0.0337
	(0.0095)	(0.0148)	(0.0066)	(0.0105)	(0.0260)
Recent Positive Reviews	0.0001*	0.0003*	0.0002*	0.0016*	0.0072*
	(0.0000)	(0.0001)	(0.0001)	(0.0005)	(0.0012)
Recent Negative Reviews	-0.0000	-0.0024*	-0.0049*	-0.0109*	-0.0027
	(0.0000)	(0.0003)	(0.0007)	(0.0018)	(0.0099)
Recent Sentiment	0.0005	0.0053	0.0098*	0.0178*	0.0087*
	(0.0019)	(0.0030)	(0.0019)	(0.0026)	(0.0042)
Cumulative Sentiment	0.0116*	0.0154*	0.0143*	0.0101*	0.0034
	(0.0023)	(0.0047)	(0.0029)	(0.0040)	(0.0066)
Twitch Viewers	0.0297*	0.1520*	0.0302*	0.1076*	0.0636
	(0.0031)	(0.0216)	(0.0121)	(0.0338)	(0.2129)
Twitch Channels	-0.0008*	0.0012	0.0047*	0.0103	0.3830*
	(0.0001)	(0.0019)	(0.0009)	(0.0078)	(0.1675)
YouTube Views	0.0067*	-0.0111	0.0086*	-0.0233*	0.0108
	(0.0002)	(0.0071)	(0.0013)	(0.0065)	(0.0262)
YouTube Videos	0.0253*	0.0439*	0.0291*	0.0436*	0.0217
	(0.0020)	(0.0034)	(0.0020)	(0.0045)	(0.0113)
Users	1,461	1,461	1,461	1,461	1,461
Games	112	102	314	310	233
Observations	10,541,431	9,523,364	32,182,647	32,694,057	19,616,970

Table continued on the following pages.

Prices in each column denote the right-endpoint of each price band.

Log-odds estimates reported, coefficients marked (*) are statistically significant at the 5% level.

Table 2.5: Individual Logit Demand Estimates - Continued

	Price Band				
Covariate on $\Pr(q_{ijt} = 1)$	\$44.99	\$29.99	\$14.99	\$4.99	\$0.99
Single-Player	0.6690	-0.3885	-0.2268	-0.0651	-0.3678
	(0.5045)	(0.3467)	(0.2345)	(0.5092)	(0.2566)
Multi-Player	0.1499	0.4190*	0.1986	0.5781*	-0.3656
	(0.0796)	(0.1294)	(0.1051)	(0.1424)	(0.2784)
Cooperative	0.0634	0.2744*	0.1326	-0.6662*	-0.1526
	(0.0536)	(0.1329)	(0.0815)	(0.1487)	(0.3872)
Massively Multi-Player	-0.8928*	-0.6321*	-0.3282*	-0.0646	0.0032
	(0.4228)	(0.2821)	(0.0571)	(0.1018)	(0.0047)
Virtual Reality	1.0048*	0.5656*	0.2073	-0.0084	0.5571
	(0.1411)	(0.2104)	(0.2264)	(0.3855)	(1.1730)
Trading Cards	-0.3247*	0.2597*	-0.3475*	0.2274	0.5855
	(0.0597)	(0.1114)	(0.0744)	(0.1301)	(0.3071)
Tracked Stats	-0.3337*	-0.0173	0.1743*	0.4079*	0.1711
	(0.0516)	(0.0929)	(0.0474)	(0.0786)	(0.1465)
Mod Support	0.3622*	0.6255*	0.1821*	0.8965*	-0.7765
	(0.0605)	(0.1088)	(0.0823)	(0.1122)	(1.0156)
Has Achievements	0.1235	0.0687	0.1531	0.0369	0.0521
	(0.0912)	(0.0421)	(0.2941)	(0.0259)	(0.0511)
Fixed Effects	Game j and Week t				
Users	1,461	1,461	1,461	1,461	1,461
Games	112	102	314	310	233
Observations	10,541,431	9,523,364	32,182,647	32,694,057	19,616,970

Prices in each column denote the right-endpoint of each price band.
Estimates reported as log-odds units.
Coefficients marked (*) are statistically significant at the 5% level.

bands as well. Also consistent with the aggregate-level findings are the similar estimates for the effects of ownership share, age, and median hours played. The estimated demand effects of a larger number of supported languages is more economically intuitive in the findings of the individual-level model with significant positive coefficients at higher price bands.

Table 2.5: Individual Logit Demand Estimates - Continued

	Price Band				
Covariate on $\Pr(q_{ijt} = 1)$	\$44.99	\$29.99	\$14.99	\$4.99	\$0.99
Network Degree	0.0012*	0.0030*	0.0022*	0.0031*	0.0024
	(0.0004)	(0.0008)	(0.0005)	(0.0007)	(0.0014)
Second Degree	-0.0001*	-0.0001*	-0.0001*	-0.0001*	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Network Support	0.6032*	0.4947*	0.2652	0.4443	0.6866
	(0.1176)	(0.2310)	(0.1623)	(0.2486)	(0.4675)
Network Demand	1.5481*	1.4876*	2.3737*	2.9093*	2.9777*
	(0.1698)	(0.2929)	(0.1927)	(0.4289)	(0.7363)
Network Ownership	0.0226*	0.0652*	0.0592*	0.0367*	0.1662
	(0.0032)	(0.0166)	(0.0121)	(0.0122)	(0.0947)
Supported Demand	1.4065*	1.1773*	1.9518*	2.2182*	1.6902*
	(0.1696)	(0.2896)	(0.1907)	(0.4279)	(0.7651)
Supported Ownership	1.3762*	1.2419*	1.2279*	2.2898*	1.0949*
	(0.0545)	(0.1325)	(0.0910)	(0.1370)	(0.5208)
Users	1,461	1,461	1,461	1,461	1,461
Games	112	102	314	310	233
Observations	10,541,431	9,523,364	32,182,647	32,694,057	19,616,970

Prices in each column denote the right-endpoint of each price band.
Estimates reported as log-odds units.
Coefficients marked (*) are statistically significant at the 5% level.

Interestingly, the estimated influence of aggregate measures of community sentiment are not dampened at the individual level when adding local network measures of sentiment to the model. This suggests there are multiple complementary layers to social influence on demand, that of the overall community as well as that of one's immediate peer group. These effects may be complementary or countervailing, however the inclusion of network measures of peer demand does not absorb the estimated impact of aggregate community sentiment. The estimated demand elasticities for user reviews, Twitch viewers, and YouTube videos are

stable across price bands.

In contrast to the rich set of features included in the aggregate model, the individual level specification affords us the opportunity to include a large number of game j and time period t fixed effects which control effectively for the diverse set of game-specific features and their influence on realized demand. We do still include a number of key features which are common to the set of products we study. The effects of these characteristics are presented in the 2nd part of Table 2.5. For the most part, the directional impacts of these characteristics are consistent with those reported previously.

The most interesting set of results regarding the impact of peer influence on individual demand outcomes are reported in the 3rd section of Table 2.5. The coefficients on network degree estimate the incremental probability of purchase for each additional friend in user i's active peer network. These estimates suggest that more socially active users on Steam are also more likely to purchase games across all price bands. The coefficients on second degree are fascinating; reporting a significantly negative effect. This is a somewhat counter-intuitive finding which suggests that growth in the number of friends-of-friends in an individual's network may diminish their resulting demand. A theory which could support this finding is a *paradox of choice* by which a second degree social network which is too large presents an individual with a lack of clarity in peer recommendation which is preventative towards decisive action. The estimated effects surrounding network support, however, are quite unambiguous. These coefficients clearly suggest that strengthening the support of a social network by adding more mutual friends increases the probability of purchasing games on Steam, particularly within the upper-most price bands.

Lastly, I examine the interaction between social network structure and demand outcomes and present a set of coefficient estimates which represent the primary contribution of this paper.

I estimate that for each additional peer who purchases game j in week $t-1$ the probability of user i purchasing the same game in week t more than doubles. The magnitude of this effect on the log-odds of purchase is more than twenty times stronger than the estimated effect of a 10% discount in the sales price of the game. Moreover, the estimates on the effect of supported demand imply that if the connected peer who purchases game j is supported by a mutual friend who also owns game j the incremental likelihood of purchase more than doubles again. While the largest estimated effects are observed for prior-period demand, we also observe significant positive effects from network ownership and supported ownership which suggests that as more peers in user i's social network own game j the likelihood of that user purchasing game j increases rapidly.

2.5.3 Robustness of Individual Demand

A potential concern with the measures of social influence in the logit model from section 2.5.2 is the possibility that the findings are driven by omitted variable bias due to unobserved game-specific or time-varying characteristics which are not included in the model. As a robustness check for these estimates, I also estimate a linear probability model which controls for game-time effects through a large number of nuisance parameters γ_{jt}.

$$q_{ijt} = \beta_D' D_{it} + \beta_S' S_{ijt} + \gamma_{jt} + \epsilon_{ijt} \tag{2.6}$$

The nuisance parameters γ_{jt} replace explicit control for prices and game-specific characteristics, allowing for identification of the influence of social effects S_{ijt} using residual demand relative to the game-time specific control.

Table 2.6 presents results from this specification after controlling for γ_{jt} effects. These coefficients reveal a comparable story to that presented by the logistic model, however several

Table 2.6: Linear Probability Estimates

	Price Band				
Covariate on q_{ijt}	\$44.99	\$29.99	\$14.99	\$4.99	\$0.99
Network Degree	0.0025*	0.0022*	0.0013*	0.0009*	0.0002
	(0.0009)	(0.0005)	(0.0002)	(0.0001)	(0.0001)
Second Degree	-0.0001*	-0.0001*	-0.0000*	-0.0000*	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Network Support	0.4746*	0.3329*	0.1079*	0.1036*	-0.0158
	(0.1116)	(0.0559)	(0.0252)	(0.0162)	(0.0132)
Network Demand	8.4680*	7.2608*	6.9385*	5.0629*	2.5046*
	(1.2419)	(0.7265)	(0.3857)	(0.3960)	(0.3757)
Network Ownership	-0.0159	0.2789*	0.0822*	0.1350*	0.1193*
	(0.0200)	(0.0401)	(0.0200)	(0.0205)	(0.0390)
Supported Demand	24.5095*	22.6239*	26.9580*	29.9565*	16.9298*
	(1.2550)	(0.8112)	(0.4282)	(0.4418)	(0.5486)
Supported Ownership	1.4836*	-0.2002	0.2858*	1.0666*	0.0969
	(0.1641)	(0.1734)	(0.0869)	(0.0954)	(0.1532)
Game-Time Effects γ_{ijt}	Yes	Yes	Yes	Yes	Yes
Mean Probability q_{ijt}	9.78e-10	2.56e-09	6.42e-10	2.61e-10	2.90e-11
Users	1,461	1,461	1,461	1,461	1,461
Games	112	102	314	310	233
Observations	10,541,431	9,523,364	32,182,647	32,694,057	19,616,970

Prices in each column denote the right-endpoint of each price band.
Coefficient estimates presented in basis points for ease of interpretation.
Coefficients marked (*) are statistically significant at the 5% level.

price bands and social effects vary in their relative magnitude and significance. The strong findings regarding the influence of network support and peer demand are preserved under this specification, providing reassurance that the logistic model was not overly affected by the risk of omitted variable bias.

2.6 Conclusion

In this paper, I demonstrate the effectiveness of peer influence and social network richness in motivating individual demand outcomes on the Steam platform using a novel dataset that effectively pairs observable demand outcomes with dynamic measures of socialization. I contrast aggregate-level demand estimates with those from a granular individual-level model and demonstrate that the measured network and peer effects appear to be incremental to the contribution of previously estimated factors. This motivates two recommendations for further study of such markets. Firstly, that identifying time-varying measures of consumer heterogeneity, particularly in cases where social structures and peer influence may have a powerful effect, is critical to establish a thorough understanding of individual consumer behavior. Secondly, from a market design perspective, that the developers of socially networked markets can benefit from carefully understanding the role that socialization plays in determining market outcomes. Such understanding has the potential to increase social welfare by inspiring more effective market designs and allowing customers to capture greater surplus at lower cost.

Chapter 3

THE PREDICTIVE POWER OF INVESTOR SENTIMENT: EVIDENCE FROM THE CAPS PLATFORM

3.1 Introduction

Investigation into the causal influence of peers on individual behavior or outcomes represents an exciting and active field of research for applied econometricians. Recent work in the market for financial securities has shown that the investment decisions made by individual investors are affected by the portfolio composition and outcomes of friends, neighbors, or colleagues. In a high-stakes field experiment, Burstzyn et al. (2014) demonstrated that the decisions made by institutional investors at a major brokerage were significantly influenced by social learning. Investors who observed peers actively trading in certain securities were incrementally likely to also invest. In Chapter 1, I characterize the behavior of household investors on the Motley Fool platform who issue directional predictions on forward returns and observe the predictions made by their socially networked peers, showing that individuals are strongly motivated by the behavior of their peer group. These empirical findings imply that investors and analysts perceive predictive value in the observed choices of others, a hypothesis which contradicts traditional models of efficient markets in which all relevant information is synthesized into an asset's traded price.

This paper investigates the merit of an alternative *wisdom of crowds* hypothesis which considers whether the aggregated insights of many individuals, each of whom possesses some specialized information, improves the predictability of future asset returns. A standard

narrative for why crowd wisdom might accumulate involves the possibility that certain individuals are sophisticated in their ability to predict returns; the choices of these sophisticated investors may be under-weighted by the market mechanism which informs price.[1] Alternatively, an ensemble of disparate predictions generated through a variety of model assumptions can generate a combined forecast through optimal prediction pooling.[2]

Evidence in favor of crowd wisdom in financial markets is provided by Chen et al. (2011) who extract measures of sentiment from articles and comments posted on the SeekingAlpha website and found a significant relationship between the directionality of sentiment and future returns for that asset. A potential hazard in the presence of reactivity to peer behavior is the risk that sentimental herding behavior may occur and generate systematic market inefficiencies if sentiment which motivates investor behavior is biased or uninformative in predicting future returns.[3] An example is shown in recent research by Jannati et al. (2016) who identify significant systematic biases in the earnings forecasts made by stock market analysts. These biases are attributed to in-group effects, where analysts are more likely to give a high rating to companies whose CEO shares demographic or ideological traits with the analyst.

The CAPS platform on Motley Fool is a social investing platform where users submit predictions regarding the expected performance of publicly-traded assets relative to the returns offered by the S&P500 index.[4] CAPS was created with the *Wisdom of Crowds* hypothesis in mind, encouraging users to contribute their own opinions and insights regarding the future

[1]Barbaris, Shleifer, and Vishny (1998) articulate a simple model of this behavior with respect to print media prior to the arrival of internet investment platforms.

[2]Geweke (2011) provides theory for how a diversity of predictions generates improvements in forecast accuracy through optimal pooling and provides an applied example of its benefits for prediction of S&P500 returns.

[3]See Shiller (2000) for a historical overview of sentimental herding amongst individual investors.

[4]The CAPS platform is available at http://caps.fool.com

performance of a stock. The collective set of these predictions informs the community rating each asset. A cornerstone motivation for the CAPS platform is to test the efficient markets hypothesis against the alternative that the combined wisdom of individuals each possessing limited information can accurately assess future stock prices. CAPS players compete for community prestige awarded to those players who demonstrate a history of successful prediction. The efficacy of aggregate CAPS sentiment has been examined by Avery, Chevalier, and Zeckhauser (2016) who consider whether the predictions made by CAPS users can aggregate to provide meaningful guidance regarding future stock prices.[5]

The authors test this hypothesis by constructing a simulated portfolio using a heuristic strategy based on quintiles of proportional sentiment and apply a Fama-French factor decomposition to attribute returns from this sentiment-based portfolio to loadings on market-level risk factors. The authors conclude that the aggregated predictions of CAPS users can outperform the market and generate risk-adjusted excess returns attributable to a stock picking effect.

> Top quintile stocks are estimated to outperform the model by 29 basis points per trading day, whereas bottom quintile stocks are estimated to underperform the model by 21 basis points per trading day. This difference in returns of 50 basis points per day corresponds to a 12.8% difference in returns on an annualized basis (or 13.7% with compounded interest). Thus, known factors such as momentum cannot fully explain our results.

A key limitation of the Avery, Chevalier, and Zeckhauser analysis is the lack of quantification surrounding the gains from aggregate sentiment for predictive power relative to a comparable baseline approach. This may cause the impact of crowd wisdom to be over-stated. In contrast, the definition of proportional sentiment considered is relatively simplistic, neglecting some richness of data which helps to contextualize the value of each prediction. More

[5] An initial draft of this paper was published as a 2011 NBER working paper, and later formally published in the *Review of Finance*.

sophisticated measures of aggregate sentiment may extract further value from the data.

To address remaining questions in understanding the role of investor sentiment in the CAPS context, we apply a more conventional approach towards examining the predictive contribution of CAPS platform sentiment by employing a formal test for predictive power across over 5,000 distinct assets and contrasting the returns generated by two alternative portfolios. We consider excess returns and aggregate investor sentiment as a multivariate system with a possible long-run cointegrating relationship and estimate tests for Granger causality which follow the standard Toda-Yamamoto (1995) procedure. This test robustly identifies using an augmented VAR model whether aggregate sentiment can improve predictability of forward returns at the individual asset level.[6] To supplement the findings of the Toda-Yamomoto tests, we estimate expected return and conditional volatility using a GARCH framework. We characterize the gains for prediction from the inclusion of aggregate sentiment. We contrast the performance of a sentiment-aware model specification with a myopic baseline and demonstrate evidence of excess returns not attributable to common market risk factors.

3.2 The CAPS Platform

This paper builds upon my own work investigating the role of peer influence on the predictions issued by CAPS market participants which uses a panel dataset of CAPS user activity paired with asset price and characteristic data.[7] The core data component is the complete prediction history of 892 socially connected users on the CAPS platform, generating a panel dataset containing 634,178 asset predictions from June 2006 through December 2014. Selected assets j are paired with prices P_{jt} from Yahoo Finance which are adjusted for stock splits, dividends

[6]Toda and Yamamoto (1995) and later Dolado and Lütkepohl (1996) show that a modified Wald test applied to the lag coefficients from an augmented VAR model is asymptotically efficient in identifying predictive relationships between series, of which some may be non-stationary.

[7]See Clayton (2017 Working Paper).

disbursements, and other redistributions. We define the asset level return R^h_{jt} at horizon h as the return from holding asset j from period $t-h$ until period t.[8]

$$R^h_{jt} = \frac{(P_{jt} - P_{jt-1})}{P_{jt-h}} \quad \text{and} \quad r^h_{jt} = \ln\left(\frac{P_{jt}}{P_{jt-h}}\right) \tag{3.1}$$

Similarly, we define market-level returns R^h_{Mt} as the horizon h return from holding the SPY fund which indexes the S&P500 from period $t-h$ until period t. When estimating models, we will utilize log returns denoted r^h_{jt} for their convenient properties and normalized distribution.[9]

The objective of the CAPS game is to predict which assets will outperform, and which will underperform relative to the S&P500 index. Excess return on the market for each asset is defined as the difference between asset-specific returns and the S&P market return, Y_{jt}. Likewise, we define log excess returns y_{jt} as the difference in log returns.

$$Y_{jt} = R_{jt} - R_{Mt} \qquad \text{and} \qquad y_{jt} = r_{jt} - r_{Mt} \tag{3.2}$$

Players $i \in N_t$ who participate in the CAPS prediction platform each possess some asymmetric information and form heterogeneous expectations of period $t+1$ excess returns using that information:

$$E_{ijt}\left[Y_{jt+1}\right] = E_{ijt}\left[R_{jt+1} - R_{Mt+1}\right] \tag{3.3}$$

Given their individual assessment of each asset's expected return, players select strategies

---

[8] This paper focuses largely on 1-step ahead returns. For convenience in notation we allow the absence of a superscript h denote a one-step-ahead prediction where $h = 1$. Where possible, we draw notation from John Campbell's textbook "The Econometrics of Financial Markets".

[9] While the standard excess return is the objective evaluation metric on the CAPS platform we use log-returns for the purposes of econometric estimation where we assume $r_{jt} \sim \mathcal{N}\left(\mu_j, \sigma^2_j\right)$ and the difference in log returns is normally distributed $y_{jt} \sim \mathcal{N}\left(\mu_j - \mu_M, \sigma^2_j + \sigma^2_M - 2\sigma_{jM}\right)$.

$s_{ijt} \in \{-1, 0, 1\}$ which map expected returns to discrete predictions. An outperform prediction implies that the player expects asset returns to exceed market return. Conversely, an underperform prediction conveys the expectation that asset returns will be surpassed by market-level return. These individual-specific expectations are evaluated against risk thresholds $\bar{\gamma}_i$ and $\underline{\gamma}_i$ which allow for heterogeneous investor preferences. The decision rule which results in active positions held by each CAPS user can be described as follows:

- ◇ Outperform: $s_{ijt} = 1$ if $E_{ijt}[Y_{jt+1}] \geq \bar{\gamma}_i$
- ◇ Underperform: $s_{ijt} = -1$ if $E_{ijt}[Y_{jt+1}] \leq \underline{\gamma}_i$
- ◇ No prediction: $s_{ijt} = 0$ if $\underline{\gamma}_i < E_{ijt}[Y_{jt+1}] < \bar{\gamma}_i$

Players issue a prediction each period, either explicitly or implicitly. In the absence of an explicit prediction, the player will retain their current position from the prior trading period such that $s_{ijt} = s_{ijt-1}$. Players begin with a neutral position $s_{ijt} = 0$ for all assets and may update their positions at any time. Positions are mutually exclusive; a user may only have one active prediction for a certain stock at any point in time.

Incentives for CAPS participants are provided through a publicly visible score π_{it} which provides a reputation-based currency for investing enthusiasts on the CAPS platform.[10] Each period, players receive a contribution to their score as the sum of basis point excess return achieved by each of their predictions. The incremental score earned by player i in period t is given by:

$$\pi_{it} = 100 \cdot \sum_t \sum_j [s_{ijt-1} Y_{jt}] \tag{3.4}$$

[10] The CAPS score is visible on each user's profile as well as on a global public ranking of players. The most highly ranked player is displayed prominently on the home page as the "Top Fool".

Each player's lifetime score Π_{it} is given by the cumulative sum of prior daily scores:

$$\Pi_{it} = \sum_{k=t_0}^{t} \pi_{ik} \tag{3.5}$$

3.2.1 Characterizing Aggregate Sentiment

The set of active positions $S_{jt} = \{s_{ijt} | s_{ijt} \neq 0\}$ held by players for asset j in period t forms the basis for aggregate measures of CAPS investor sentiment. We consider the definition of proportional sentiment proposed by Avery, Chevalier, and Zeckhauser which defines sentiment as the share of currently active predictions which predict positive excess return. We denote the proportional sentiment index as:

$$X_{jt}^{P} = \frac{\sum_i 1\{s_{ijt} = 1\}}{\sum_i 1\{s_{ijt} \neq 0\}} \tag{3.6}$$

We consider an improvement upon proportional sentiment in which the contribution of each observed prediction is weighted by $\hat{w}_{ijt}$, which describes the expected likelihood of that prediction being directionally accurate. We define directional accuracy $w_{ijt} \in \{0, 1\}$ as:

$$w_{ijt} = 1\{s_{ijt} Y_{jt+1} > 0\} \tag{3.7}$$

We can estimate the expected probability of a directionally-accurate prediction using a rich set of covariates Z_{ijt} which describe observable asset-specific and player-specific characteristics.

$$\Pr\left(w_{ijt} | s_{ijt}, Z_{ijt}\right) = \frac{\exp\left(Z'_{ijt}\theta\right)}{1 + \exp\left(Z'_{ijt}\theta\right)} \tag{3.8}$$

We use a logistic regression framework to estimate the likelihood of observing an outcome consistent with the submitted prediction conditional on the history of observed features.

The set of possible features Z_{ijt} is reasonably large; thus we employ dual-regularization to perform feature selection and mitigate the risk of over-fitting our predicted weights $\hat{w}_{ijt}$. We estimate the above model using penalized maximum likelihood which may apply both L1 and L2 regularization to the conditional log-likelihood function. The regularized conditional log-likelihood is given by:

$$\ell(w) = \ln \prod \Pr\left(w_{ijt}|s_{ijt}, Z_{it}, \theta\right) - \left[\omega\lambda||\theta||_1\right] - \left[(1-\omega)\frac{\lambda}{2}||\theta||_2^2\right] \tag{3.9}$$

We estimate the parameters θ by stochastic gradient descent and select the penalization parameter λ using cross-validation against a reserved sample of test users. The penalty applied to the L1 norm (ω) eliminates unhelpful features from the model while the penalty applied to the L2 norm ($1-\omega$) helps to regularize the influence of included features and reduce the risk of over-fitting. We calibrate the L1 weight ω in order to achieve a desired density of included features.[11] Appendix .5.1 details the set of possible features Z_{ijt} which were considered. In total, the model considers 383 potential features including a rich set of fixed effects. The regularized logistic model selects 76 features for inclusion in the final model. The predicted probabilities $\hat{w}_{ijt}$ generated by this model are used directly as weights to construct a weighted sentiment index X_{jt}^W for each asset.

$$X_{jt}^W = \frac{\sum_i \left[\hat{\pi}_{ijt} 1\{s_{ijt}=1\} + (1-\hat{\pi}_{ijt}) 1\{s_{ijt}=-1\}\right]}{\sum_i 1\{s_{ijt} \neq 0\}} \tag{3.10}$$

In general the features considered for inclusion in the weight model are drawn from four categories of covariates, following the structure of my related work using CAPS data. The first group of covariates measures asset characteristics including historical performance, trad-

[11] See Zou and Hastie (2005) for the seminal paper on dually regularized linear optimization methods. Advice for practitioners in extending this framework to generalized linear models and estimating parameters via coordinate descent is detailed in Friedman et al. (2010).

ing volume, a variety of measures from financial reporting, and a set of industry and sector classifiers.[12] The second group of covariates includes macroeconomic features including Fama-French factors and industry-level returns.[13] The third set of covariates captures elements of individual heterogeneity specific to the predicting player. These heterogeneous characteristics include CAPS platform tenure, historical portfolio performance, measures of industry or sector emphasis, and quantifications of social engagement. Lastly we include prediction-specific information measuring freshness of each active position, whether it was accompanied by a corresponding text-based pitch, and the historical success of the predicting player within the same industry and market sector.[14]

Figure 3.1 depicts the observed relationship between excess returns Y_{jt}, proportional sentiment X_{jt}^{P}, and weighted sentiment X_{jt}^{W} for General Electric between 2011 and 2012. Percentage point excess return is labeled on the left-hand y-axis, while sentiment indices are labeled on the right. Observations where proportional sentiment exceeds 0.5 denote periods where the majority of observed CAPS players expect the asset to outperform the S&P500 index in future periods, attaining positive excess return. Larger values of weighted sentiment correspond to periods with a greater expected probability of outperform predictions being directionally accurate. For many assets the weighted index differs in both magnitude and variation with it's proportional counterpart, as demonstrated in the case of GE.

In the following sections we build upon these measures of aggregate sentiment to provide formal tests for their contribution towards predictive power in section 3.3. In section 3.4 we incorporate these sentiment measures as covariates in a GARCH model for expected mean and conditional variance. Lastly in section 3.5 we perform a rich portfolio simulation in

[12]These covariates are drawn from Yahoo Finance and the CRSP/CompuStat merged database.

[13]Macroeconomic series are obtained from Kenneth French's personal website and CRSP/CompuStat.

[14]Individual-specific and prediction-specific measures are drawn from collected CAPS platform data.

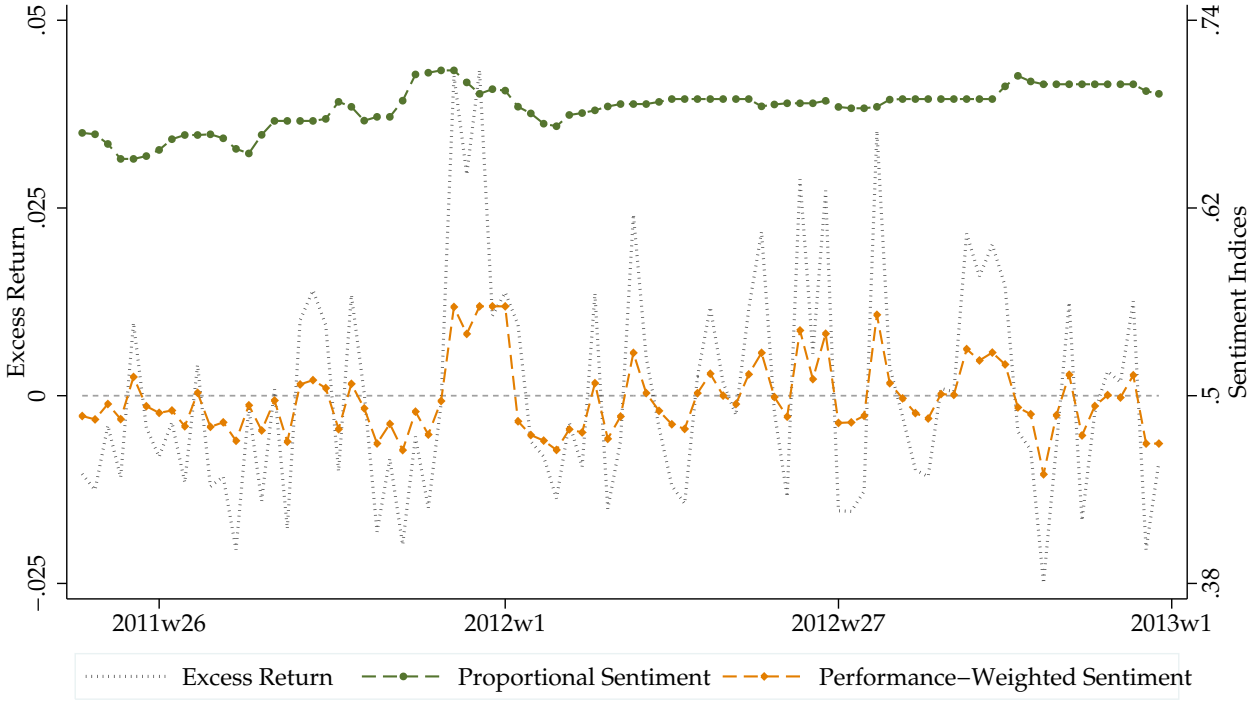

Figure 3.1: General Electric (GE) Excess Returns and Aggregate Sentiment

which we contrast the performance of the myopic GARCH models with the performance of their sentiment-aware counterparts and estimate by Fama-French factor decomposition the excess returns generated by each simulated portfolio.

3.3 Testing for Predictive Power

Our goal is to measure the frequency and degree to which measures of aggregate sentiment can improve the predictability of future excess returns. We employ standard methods to test aggregate sentiment for Granger causality which is defined as follows; in the case of two time series X_t and Y_t, X is said to Granger-cause Y if Y can be better predicted using the histories of both X and Y than it can by using the history of Y alone.[15] In our context, Y_{jt}

[15]See Granger (1988) for the seminal paper on testing for predictive causality.

is the excess return on the market, y_{jt} is the log-excess return which we use for estimation, and X_{jt}^P and X_{jt}^W are our candidate predictors.

The standard Granger causality testing procedure involves estimating a vector auto-regression on X and Y and testing for the joint significance of the coefficients β_k which modify the lagged terms of X_{jt}. This methodology was improved upon by Toda and Yamamoto (1995) and later Dolado and Lütkepohl (1996) to produce an unbiased test for series which may be non-stationary or have a cointegrating relationship. The Toda-Yamomoto procedure relies on careful selection of the VAR order P, which is selected as the lag order which maximizes information criteria for the model with order K, plus the maximum order of integration M over the set of time series.[16] The additional auto-regressive order is included in the model to allow a modified Wald test statistic to feature an asymptotic χ^2 distribution even in the face of cointegration. We apply the Toda-Yamomoto approach to measures of log-excess return and CAPS investor sentiment, estimating the following vector auto-regression with order P as:

$$\begin{aligned} y_{jt} &= \alpha_0 + \sum_{k=1}^{P} \alpha_k y_{jt-k} + \sum_{k=1}^{P} \beta_k X_{jt-k} + u_{jt} \\ X_{jt} &= c_0 + \sum_{k=1}^{P} c_k X_{jt-k} + \sum_{k=1}^{P} d_k y_{jt-k} + v_{jt} \end{aligned} \tag{3.11}$$

Using the estimated coefficients β_k on the lags of sentiment in the excess return equation, we use a Wald test for the joint significance of sentiment against the null hypothesis that

[16] We use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for stationarity. In practice, y_{jt} is almost always stationary (by definition) while aggregate sentiment X_{jt} may often be non-stationary for certain assets. As a result, the VAR system typically has a maximum integration order of $M = 1$.

sentiment provides no information for future excess returns.

$$H_0 : \beta_1 = ... = \beta_k = 0, \quad W_k \xrightarrow{D} \chi^2 \tag{3.12}$$

This procedure involves the testing of multiple hypotheses, so we must account for the family-wise error rate (or false discovery rate) implicit in our findings. We report results using the Benjamini-Hochberg (B-H) step-up procedure which constrains the number of claimed rejections using the family-wise error rate.[17] Table 3.1 reports the proportion of the 4,880 testable time series which report significant Wald statistics against the χ^2 distribution.

Table 3.1: Toda-Yamomoto Tests for Granger Causality

		Proportional Sentiment		Weighted Sentiment	
Procedure	Assets	$p^{TY} \leq 0.10$	Pct.	$p^{TY} \leq 0.10$	Pct.
Unadjusted	4,880	597	12.23%	941	19.28%
Hochberg Step-Up	4,880	25	0.05%	42	0.09%

Each entry reports the likelihood of rejecting the T-Y null hypothesis.
H_0: X_{jt} is not Granger causal for forward excess returns Y_{jt}.

Without applying the appropriate family-wise error correction we would reject the null hypothesis that proportional sentiment has no effect on forward returns for 12.23% of assets while rejecting 19.28% of assets using our measure of weighted sentiment. As this is a multiple hypothesis testing problem, the collection of rejections is subject to inherent type 1 error which we attempt to control for using the B-H procedure which dramatically reduces the

[17] In particular, the Benjamini-Hochberg procedure involves limiting the number of null hypothesis rejections by inflating the requisite critical value for each test as the number of tests increases in order to account for the family-wise error rate. For m hypotheses and a target significance level α, we order the hypotheses in ascending order of their estimated p-values and find the largest k such that $P_k <= \frac{k}{m}\alpha$. We then recognize a rejection of the null hypothesis for tests 1 through k while failing to reject the remaining hypotheses. This procedure constrains the collective set of rejections to account for the inherent likelihood of type 1 error across a large number of independent tests.

number of rejections which may be safely claimed.[18] Under the B-H correction we may only safely claim rejections for 25 assets using proportional sentiment and for 42 assets with the weighted sentiment. This procedure is highly penalizing particularly in the case of a large number of hypotheses. In both cases, we observe that estimation of heterogeneous prediction weights increases the likelihood that sentiment will be Granger causal predictor of forward excess returns, but that this sentiment index is likely only meaningful for the prediction in the case of a small number of assets.

3.4 Forecasting Returns

The Granger causality results suggest that gains to predictability of excess returns are found only for a limited subset of assets. To complement the in-sample approach of the Toda-Yamomoto test, we consider an alternative approach towards evaluating the benefit of these sentiment measures for prediction by including them as covariates in a predictive model and evaluating the out-of-sample performance of various model specifications. We produce one-step-ahead forecasts of log-excess return y_{jt} and volatility σ^2_{jt} using a generalized auto-regressive conditional heteroskedasticity (GARCH) model to quantify the benefits of including sentiment indices for both predictive accuracy and simulated portfolio returns. We estimate a standard GARCH model where the mean equation is estimated as a low order auto-regressive moving-average process, denoted as $ARMA(P, Q)$. We augment this myopic baseline to estimate a sentiment-aware specification as an ARMA-X process which includes the level and change in X_{jt-1} in the mean equation. For both specifications we estimate the conditional variance using a fixed-order $GARCH(1, 1)$ process. Specifically, for each asset

[18]Typical family-wise error rate procedures including the B-H step-up procedure are conservative, recognizing the fewest number of hypotheses which may be safely rejected under the presence of type 1 error.

we estimated weekly log-excess returns y_t and their conditional variance σ_t^2 as:

Myopic Model:

$$\begin{aligned} y_t &= \sum_{p=1}^{P} \phi_p^0 y_{t-p} + \sum_{q=1}^{Q} \theta_q^0 \epsilon_{t-q} + \epsilon_t \\ \sigma_t^2 &= a_0^0 + a_1^0 \epsilon_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{aligned} \tag{3.13}$$

Sentiment-Aware Model:

$$\begin{aligned} y_t &= \sum_{p=1}^{P} \phi_p^1 y_{t-p} + \sum_{q=1}^{Q} \theta_q^1 \eta_{t-q} + \beta_1 X_{t-1} + \beta_2 \Delta X_{t-1} + \eta_t \\ \sigma_t^2 &= a_0^1 + a_1^1 \eta_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{aligned}$$

We estimate these two models for every asset j across every week t in an ensemble of rolling backtests to evaluate the difference in accuracy of predicted $\hat{y}_{jt}$ between the two specifications. We choose the order parameters $P <= 6$ and $Q <= 1$ independently for each asset to minimize the Bayesian Information Criterion (BIC) within the training sample. Figure 3.2 depicts an example of the resulting one-step-ahead forecasts produced for General Electric between 2011 and 2012. The GARCH predictions model the heteroskedastic variance which occurs during periods of heightened uncertainty, like in late 2011. In the case of GE, both myopic and sentiment-aware models generate similar predictions of both mean and variance.

3.4.1 Hypothesis Tests

The structure of the CAPS game suggests the *wisdom of crowds* hypothesis, $H_0 : \beta_1 > 0$ and $\beta_2 > 0$; higher levels of sentiment or upward revisions in sentiment should, on average, predict positive excess returns in subsequent periods. Furthermore, we would further expect that for series which rejected the Toda-Yamomoto hypothesis in the prior section, the β

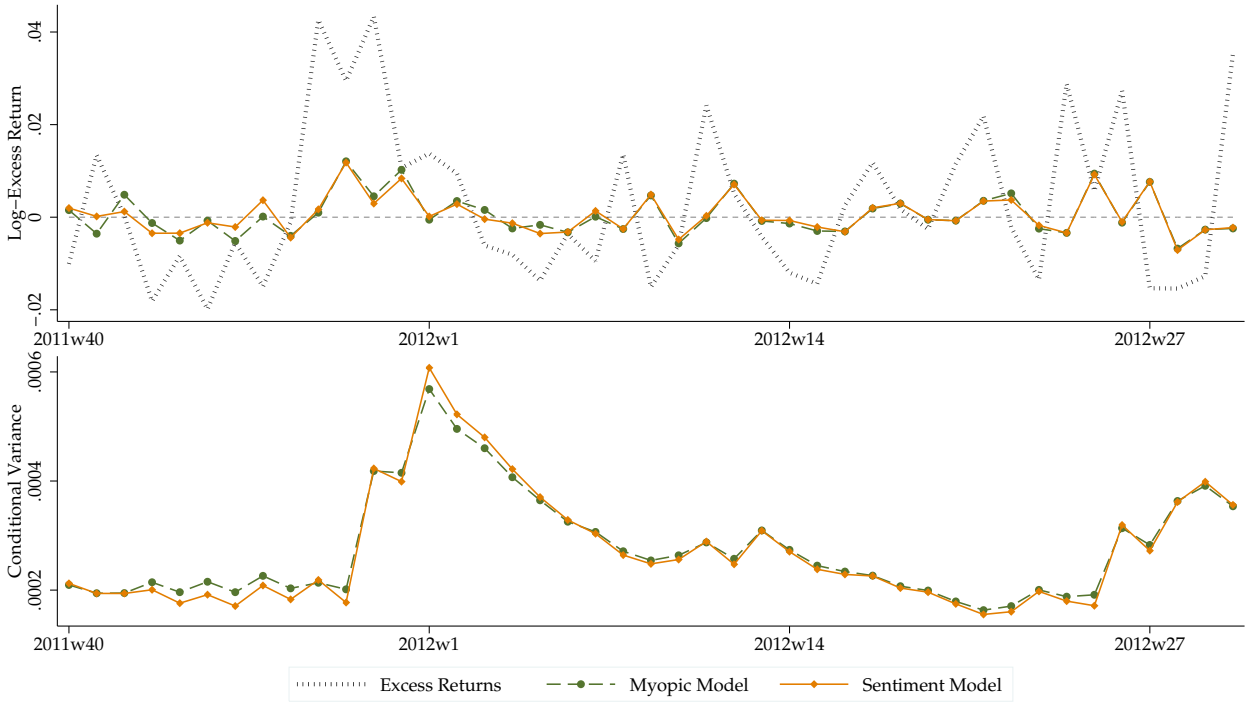

Figure 3.2: Predicted Excess Return and Volatility - General Electric

parameters should be statistically significant. We test these hypotheses for each of 5,036 distinct assets every week over the course of the training period from 2011 through 2012. These test results are reported in Table 3.2.

Table 3.2: Joint Significance of GARCH β_1 and β_2

T-Y Significant	Obs.	$p <= 0.10$	Pct.
No	388,700	114,288	29.40%
Yes	52,600	27,352	56.15%
Total	441,308	141,640	32.09%

Only 12.23% of assets were identified as Granger causal under the (unadjusted) Toda-Yamomoto procedure (Table 3.1). A Wald test of joint significance on the sentiment indices

included in the GARCH model suggests that, in total, 32.09% of assets feature variation in forward excess returns that are meaningfully explained by variation in sentiment. When examining the interaction between both hypothesis tests we observe that, for the subset of assets which rejected the T-Y null hypothesis, 56.15% also rejected a joint hypothesis on the significance of β_1 and β_2. In contrast, we only observe 29.40% rejections for assets which were not T-Y significant. These tests are not formally adjusted for family-wise error rates, but serve to characterize a degree of consistency in the usefulness of sentiment for prediction.

3.4.2 Forecast Performance

In addition to testing for crowd wisdom by evaluating the estimated parameters β_1 and β_2, we can also contrast the realized accuracy between the two model specifications. For each asset j in every week t in 2012 we predict excess returns in week $t+1$ using data from weeks $t-104$ to t. We report accuracy metrics to measure the effectiveness of included sentiment in a rolling backtest.[19] We report primarily on two standard measures of model accuracy for quantitative finance predictions; mean absolute deviation (MAD) and mean directional accuracy (MDA).[20] Absolute deviation helps us benchmark the closeness of our predictions to the realized forward return while directional accuracy helps evaluate the correctness (win/loss rate) of the prediction. We present accuracy results for both myopic and sentiment-aware model specifications and partition based on results of the initial Toda-Yamomoto test for Granger causality.

We observe that the sentiment-aware model outperforms the myopic specification across all

[19] A rolling backtest entails estimating and evaluating the model for each period iteratively where predictions for $\hat{y}_{jt+1}$ are generated using data through y_t.

[20] We examined, but do not report alternative accuracy measures including RMSE, MAPE, and MSE which produce comparable results.

Table 3.3: GARCH Backtest Accuracy Comparison

		Myopic		Sentiment	
T-Y Significant	Obs.	MAD	MDA	MAD	MDA
No	388,708	0.0181	55.98%	0.0178*	58.91%*
Yes	52,600	0.0189	55.32%	0.0186*	58.77%*
Total	441,308	0.1820	55.90%	0.0179*	58.89%*

Entries marked (*) denote the better performing model across the ensemble of backtests

assets, regardless of the initial results of the Toda-Yamomoto test, although the improvement in mean directional accuracy is greater in relative terms for T-Y significant assets. Overall, this represents a somewhat counter-intuitive finding which suggests that carefully weighted aggregate sentiment may contribute positive value to prediction for a majority of assets even if statistical power is only sufficient to identify this benefit in a minority of cases.

3.5 Portfolio Comparison

As a final measure of the contribution of aggregate sentiment towards the predictability of returns, we simulate the performance of alternative portfolios constructed using the rolling ensembles of GARCH forecasts from the previous section. Starting from January 2011 we generate simulated portfolios using three alternative prediction mechanisms: the myopic GARCH model, its sentiment-aware counterpart, and using proportional sentiment directly as an asset selection instrument. Our goal is not to propose optimal portfolio composition strategies, which is far beyond the scope of this analysis, but rather to provide a comparative view of performance under a common asset selection algorithm. Therefore we remain consistent with the quintile-based approach from the Avery, Chevalier, and Zeckhauser analysis.

Our two candidate GARCH models produce weekly forecasts of mean excess returns $\hat{y}_{jt}$

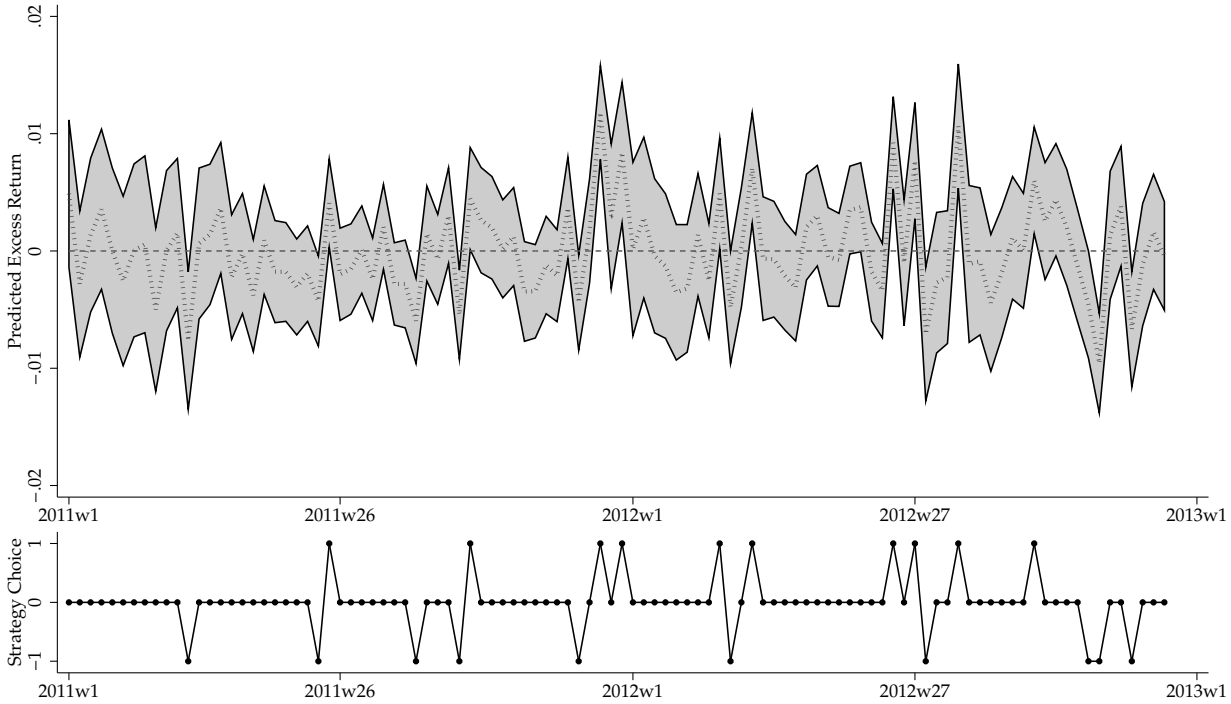

Figure 3.3: Simulated Strategies - General Electric

and conditional variance $\hat{\sigma}^2_{jt}$. We rank all 4,880 assets according to their Sharpe Ratio $q_{jt} = \hat{y}_{jt}/\hat{\sigma}_{jt}$ which allows us to select the top and bottom-most quintiles of stocks based on expected performance and volatility.[21] In the third candidate we simply select assets directly based on quintiles of proportional sentiment. Selecting the top and bottom-most quintiles of q_{jt} implies critical thresholds $(\bar{q}_{jt}, \underline{q}_{jt})$ which can be used to characterize the resulting portfolios. We use our rolling ensembles of one-step-ahead GARCH forecasts to simulate portfolio composition by selecting strategies $\hat{s}_{jt}$ as follows:

- ◇ Positive position: $\hat{s}_{jt} = 1 \quad \text{if } \hat{y}_{jt}/\hat{\sigma}^2_{jt} \geq \bar{q}_{jt}$
- ◇ Negative prediction: $\hat{s}_{jt} = -1 \quad \text{if } \hat{y}_{jt}/\hat{\sigma}^2_{jt} \leq \underline{q}_{jt}$

[21] The Sharpe Ratio is also frequently called the Information Ratio particularly in cases where returns are computed relative to some benchmark, in our case, the S&P500.

◇ No prediction: $\hat{s}_{jt} = 0 \quad \text{if } \underline{q}_{jt} < \hat{y}_{jt}/\hat{\sigma}^2_{jt} < \bar{q}_{jt}$

Figure 3.3 demonstrates how the prediction interval given by these critical thresholds generates a series of strategies $\hat{s}_{jt}$ for General Electric between 2012 and 2013. We simulate the performance of each portfolio assuming an equal stake in every chosen asset. Weekly returns for portfolio p are equal to:

$$R^p_t = \prod_j \left(1 + \hat{s}^p_{jt} r_{jt}\right) - 1 \tag{3.14}$$

The cumulative value for each portfolio is given by:

$$V^p_t = V^p_{t-1} \prod_j \left(1 + \hat{s}^p_{jt} r_{jt}\right) \tag{3.15}$$

We contrast the performance of these portfolios in Figure 3.4 which illustrate that both the sentiment-aware and myopic GARCH models outperform the S&P500 index, while sentiment-aware model generates a more successful portfolio composition. The simple quintile binning strategy on proportional sentiment performs comparably to the S&P500 index over the course of the simulation.

3.5.1 Fama-French Factor Decomposition

To quantify and contrast the advantages of the GARCH model for portfolio generation we estimate a standard Fama-French factor decomposition on portfolio returns to estimate the risk-factor loadings and excess return generated by the portfolio selection strategy, where:

$$R^p_t - rf_t = \alpha + \beta_1(R_{Mt} - rf_t) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 MOM_t + \epsilon_t \tag{3.16}$$

Results of the Fama-French decomposition are reported in Table 3.4. These results illustrate

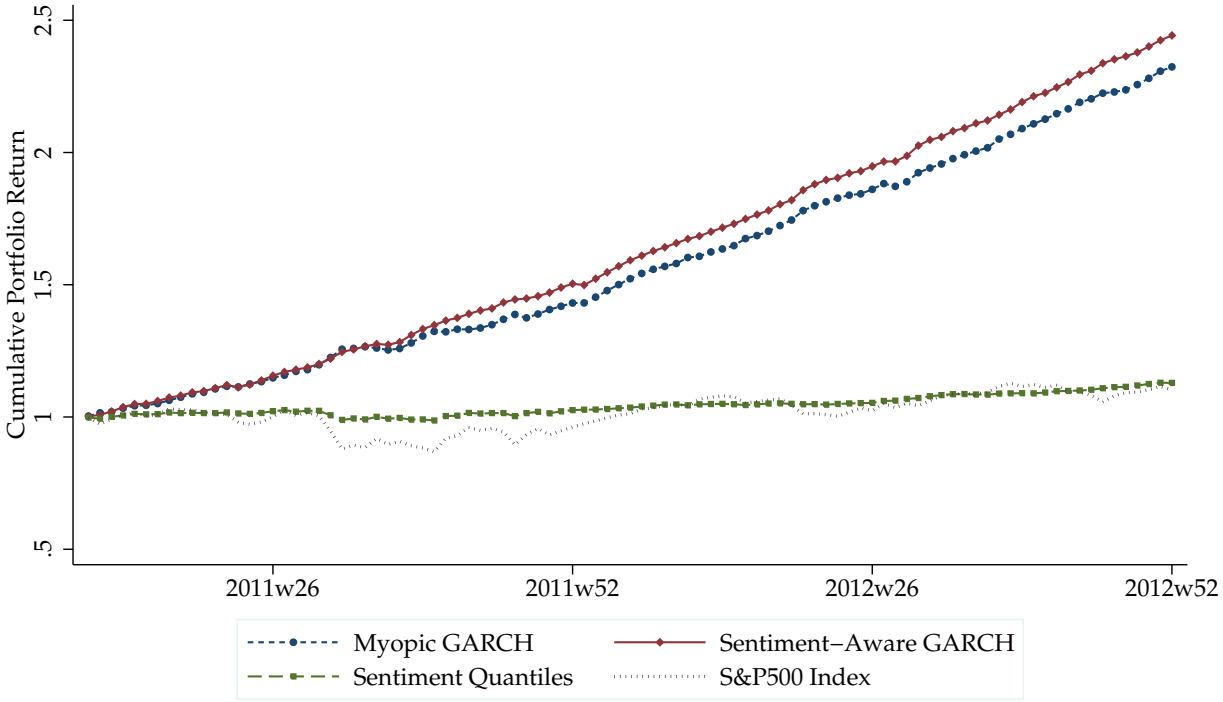

Figure 3.4: Cumulative Portfolio Returns - Model Comparison

that each of the three portfolios generates variation in returns that is explained relatively poorly by Fama-French market factors. In particular, the GARCH-based portfolios generate significant variation in returns not explained well by market factor loadings. However, the GARCH-based portfolios are estimated to produce a statistically significant risk-adjusted excess return after controlling for the portfolios correlation to market-level factors. In contrast; performance of the quintile based proportional sentiment portfolio does not generate returns which deviate meaningfully from those generated by the S&P500 index.

We test the difference in estimated parameters of excess return α between the myopic and sentiment-aware portfolios by simultaneously estimating the decomposition using seemingly unrelated regression and testing the joint hypothesis that the portfolios offer differential excess returns against the null hypothesis $H_0 : \alpha^{\text{Myopic}} = \alpha^{\text{Sentiment}}$. This test presents weak

Table 3.4: Fama-French Factor Decomposition

Risk Factor	Myopic	Sentiment	ACZ
Market Minus Risk-Free $(R_{Mt} - rf_t)$	-0.0765* (0.0272)	-0.0138 (0.0235)	0.1082* (0.0189)
Small Minus Big SMB_t	-0.0088 (0.0644)	-0.0035 (0.0557)	0.0463 (0.0448)
High Minus Low HML_t	-0.0432 (0.0725)	-0.1102 (0.0627)	0.0996 (0.0504)
Momentum MOM_t	0.0639 (0.1042)	0.0199 (0.0901)	-0.0189 (0.0724)
Excess Return α	0.0079* (0.0006)	0.0084* (0.0005)	0.0002 (0.0004)
Weeks	104	104	104
R^2	0.221	0.141	0.368

Standard errors reported in parentheses.
Estimates denoted (*) are significant at the 5% level.

evidence to reject the null hypothesis that the estimated alphas are equivalent at the 10% significance level with a χ^2 test statistic of 2.86.

3.6 Conclusion and Future Work

This paper presents moderate evidence in support of a *wisdom of crowds* hypothesis by conducting formal tests for Granger causality, evaluating the predictive contribution of aggregate sentiment in a GARCH forecasting model, and contrasting the simulated performance of portfolios. In each case, we find evidence to support the hypothesis that individual sentiment may aggregate to improve the signal and predictability of future returns over the information which is naturally captured by an asset's current price. This paper's primary contribution is to motivate further and more sophisticated evaluation of individual investor

sentiment, particularly in a real world environments, to deepen our understanding of how peer sentiment may influence investment choices and to identify whether the collective wisdom of many peers, each of whom possesses asymmetric information, can be synthesized effectively to improve market outcomes.

BIBLIOGRAPHY

[1] Angrist, Joshua D. 2014. The perils of peer effects. *Labour Economics*, 30, pp. 98-108.

[2] Avery, C. N., Chevalier, J. A., Zeckhauser, R. J. 2016. The CAPS Prediction System and Stock Market Returns. *Review of Finance*, 20.4, pp. 1363-1381.

[3] Bajari, P., Nekipelov, D., Ryan SP., and Yang M. 2015. Demand estimation with machine learning and model combination. No. w20955. *National Bureau of Economic Research*.

[4] Banerjee, Chandrasekhar, Duflo, and Jackson. 2013. "The Diffusion of Microfinance". *Science*, Vol. 341, Issue 6144.

[5] Barberis, Shleifer, and Vishny. 1998. A model of investor sentiment. *Journal of Financial Econometrics*, Vol. 49, Issue 3, pp. 307-343.

[6] Barzel, Yoram. 1982. "Measurement Cost and the Organization of Markets." *Journal of Law and Economics*, Vol. 25, No. 1, pp. 27-48.

[7] Bramoullé, Djebbari, and Fortin. 2009. Identification of peer effects through social networks. *Journal of Econometrics*, Vol. 150, Issue 1. pp 41-55.

[8] Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. "Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions". *Econometrica*, Vol. 82, No. 4, pp. 1273-1301.

[9] Campbell, John Y. and Lo, Andrew A. and MacKinlay, A. Craig. 2012. The Econometrics of Financial Markets. *Princeton University Press*.

[10] Cheng, Simon and Long, Scott J. 2007. "Testing for IIA in the Multinomial Logit Model". *Sociological Methods Research*, Vol. 35, No. 4, pp. 583-600.

[11] Chernozhukov, Victor, et al. 2016. Double machine learning for treatment and causal parameters. *arXiv:1608.00060*.

[12] Compustat. 2014. *Standard & Poor's Institutional Market Services.* McGraw-Hill Inc. http://wrds-web.wharton.upenn.edu.

[13] Corten, Rense. "Visualization of social networks in Stata using multidimensional scaling". *The Stata Journal*, Vol. 11, Number 1, pp. 52-63.

[14] Center for Research in Security Prices (CRSP). 2014. Booth School of Business. The University of Chicago. http://crsp.chicagobooth.edu.

[15] Chen, Hailiang and De, Prabuddha and Hu, Yu Jeffrey and Hwang, Byoung-Hyoun. 2013 Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, Forthcoming. Available at SSRN: https://ssrn.com/abstract=1807265

[16] Derdenger, Timothy. 2014 Technological tying and the intensity of price competition: An empirical analysis of the video game industry. *Quantitative Marketing and Economics* 12.2 (2014): 127-165.

[17] Diamond, Alexis and Sekhon, Jasjeet S. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies". *The Review of Economics and Statistics*, Vol. 95, No. 3, pp. 932-945.

[18] Dolado, Juan J. and Helmut Lütkepohl. 1996 Making Wald tests work for cointegrated VAR systems. *Econometric Reviews* 15, no. 4, pp. 369-386.

[19] Fama-French Research Returns Data. 2014. Kenneth French Data Library. http://mba.tuck.dartmouth.edu/pages/faculty/ken.french.

[20] Fama, Eugene and French, Kenneth R. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, Vol. 33, Issue 1, pp. 3-56.

[21] Fama, Eugene and French, Kenneth R. "The Capital Asset Pricing Model: Theory and Evidence." *Journal of Economic Perspectives*, Vol. 18, Issue 3, pp. 25-46.

[22] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33.1.

[23] Geweke, J. and Amisano, G. 2011. Optimal prediction pools. *Journal of Econometrics*, 164(1), pp.130-141.

[24] Grinblatt, Keloharju, and Ikäheimo. 2008. "Social Influence and Consumption: Evidence from the Automobile Purchases of Neighbors". *The Review of Economics and Statistics*, Vol. 90, No. 4, pp. 735-753.

[25] Hausman, Jerry and McFadden, Daniel. 1984. "Specification Tests for the Multinomial Logit Model". *Econometrica*, Vol. 52, No. 5, pp. 1219-1240.

[26] Hill, Shawndra and Ready-Campbell, Noah. "Expert Stock Picker: The Wisdom of (Experts in) Crowds". *International Journal of Electronic Commerce*, Vol. 15, No. 3, pp. 73-102.

[27] Hochberg, Yosef. 1988. "A sharper Bonferroni procedure for multiple tests of significance". *Biometrika*, 75, 4, pp. 800-802.

[28] Hong, H., Kubik, J.D., and Stein, J.C. 2004. "Social Interaction and Stock-Market Participation." *The Journal of Finance*, Vol. 59, pp. 137-163.

[29] Hong, H., Kubik, J.D., and Stein, J.C. 2005. "Thy Neighbor's Portfolio: Word-of-Mouth Effects in the Holdings and Trades of Money Managers." *The Journal of Finance*, Vol. 60, pp. 2801-2824.

[30] Imbens, Guido W. and Rubin, Donald B. "Causal Inference for Statistics, Social, and Biomedical Sciences". *Cambridge University Press*, April 6, 2015.

[31] Ishihara, Masakazu, and Andrew Ching. 2016. "Dynamic demand for new and used durable goods without physical depreciation: The case of japanese video games".

[32] Jackson, M. O. 2010. "An overview of social networks and economic applications". *The handbook of social economics*, 1, 511-85.

[33] Jannati, S., Kumar, A., Niessen-Ruenzi, A., and Wolfers, J. 2016 "In-Group Bias in Financial Markets". Available at *SSRN*: https://ssrn.com/abstract=2884218

[34] Kaustia, M., and Knüpfer, S. 2012. "Peer performance and stock market entry." *Journal of Financial Economics*, 104(2), pp. 321-338.

[35] King, Gary and Zeng Langche. 2001. "Logistic Regression in Rare Events Data." *Political Analysis*, Vol. 9, Issue 2, pp. 137-163.

[36] McFadden, Daniel. 1974. "Conditional logit analysis of qualitiative choice behavior." *Frontiers of Econometrics*, pp. 105-142.

[37] Manski, Charles. 2000. "Economic Analysis of Social Interactions". *Journal of Economic Perspectives*, Vol. 14, pp. 115-136.

[38] Manski, Charles. 1993. "Identification of Endogenous Social Effects: The Reflection Problem". *Review of Economic Studies*, Vol. 60, No. 3, pp. 531-542.

[39] Moretti, Enrico. 2011. "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales". *Review of Economic Studies*, Vol. 78, No. 1, pp. 356-393.

[40] Motley Fool CAPS Community. 2014. Motley Fool. http://caps.fool.com.

[41] Orland, Kyle. 2014. "Introducing Steam Gauge: Ars reveals Steam's most popular games". *Ars Technica*. https://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/

[42] Pearl, Judea. 2000. "Causality: Models, Reasoning and Inference". Vol. 29. Cambridge: MIT press.

[43] Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk". *The Journal of Finance*, Vol. 19, No. 3, pp. 425-442.

[44] Shiller, Robert J. 2000. "Irrational Exuberance." *Princeton University Press.*

[45] Shiller, Robert J., and Pound, John. 1989. "Survey evidence on diffusion of interest and information among investors." *Journal of Economic Behavior & Organization*, Vol. 12, No. 1, pp. 47-66.

[46] Sorensen, Alan. 2008. "Social learning and health plan choice." *The RAND Journal of Economics*, Vol. 37, Issue 4, pp. 929-945.

[47] Toda, Hiro Y. and Taku Yamamoto. 1995. "Statistical inference in vector autoregressions with possibly integrated processes". *Journal of Econometrics* 66, no. 1-2, pp. 225-250.

[48] Zhu, Feng, and Xiaoquan Zhang. 2006. "The influence of online consumer reviews on the demand for experience goods: The case of video games". *ICIS 2006 Proceedings*, 25.

[49] Zou, Hui and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67.2, pp. 301-320.

.1 CAPS Model and Variable Definitions

Variable	Definition
s_{ijt}	The prediction chosen where (1 = Outperform, 0 = Neutral, -1 = Underperform)
$volume_{jt}$	Millions of shares exchanged during the prior trading day
$alpha\_7_{jt}$	Estimated CAPM long-run α using rolling 7 day returns
$beta\_7_{jt}$	Estimated CAPM long-run β using rolling 7 day returns
$prior\_return\_7_{jt}$	The asset's prior rolling 7 day return
$exchange_j$	Fixed effects for the exchange on which the asset trades
$sector_j$	Fixed effects for the primary market sector in which the firm produces
$market\_return_t$	S&P500 moving average weekly return
$small\_minus\_big_t$	Fama-French SMB factor measuring differential returns of small firms relative to large ones based on market capitalization
$high\_minus\_low_t$	Fama-French HML factor measuring differential returns of high book-to-market ratio firms relative to low ratio firms
$momentum_t$	Carhart momentum factor measuring the tendency of market returns to continue moving in the same direction
$lifetime\_picks_{it}$	The cumulative number of past predictions issued by player i
$player\_tenure_{it}$	The number of months since player i first registered on the CAPS platform
$player\_up\_7_{it}$	The number of outperform predictions issued by a player over the previous 7 days
$player\_down\_7_{it}$	The number of underperform predictions issued by a player in the previous 7 days
$player_i$	Fixed effects for the CAPS player
$total\_favorites_{it}$	The number of other users marked as favorites by player i
$leader\_pick\_up_{ijt}$	The number of outperform predictions issued by higher-ranked peers during the previous 7 days
$leader\_pick\_down_{ijt}$	The number of underperform predictions issued by higher-ranked peers during the previous 7 days
$follower\_pick\_up_{ijt}$	The number of outperform predictions issued by lower-ranked peers during the previous 7 days
$follower\_pick\_down_{ijt}$	The number of underperform predictions issued by lower-ranked peers during the previous 7 days
$leader\_hold\_up_{ijt}$	The number of higher-ranked peers who actively hold an outperform position
$leader\_hold\_down_{ijt}$	The number of higher-ranked peers who actively hold an underperform position
$follower\_hold\_up_{ijt}$	The number of lower-ranked peers who actively hold an outperform position
$follower\_hold\_down_{ijt}$	The number of lower-ranked peers who actively hold an underperform position

Variable subscripts describe their level of variation in the data. Trading days are indexed by t, players are indexed by i, and securities are indexed by j.

.2 CAPS Summary Statistics

Variable	Mean	Std. Dev.	Minimum	Maximum
$volume_{jt}$	3.614	14.33	0	2,304
$alpha\_7_{jt}$	0.397	1.271	-1.801	9.679
$beta\_7_{jt}$	1.063	1.161	-1.801	3.755
$return\_7_{jt}$	0.362	6.986	-19.55	26.28
$market_t$	0.065	1.412	-6.963	6.897
$small\_big_t$	0.015	0.600	-2.015	3.566
$high\_low_t$	-0.006	0.670	-3.346	2.961
$momentum_t$	-0.051	1.196	-7.527	7.040
$total\_picks_{it}$	712.4	828.9	0	4,800
$tenure_{it}$	1,019	562.3	0	2,449
$player\_up\_7_{it}$	1.350	4.277	0	174
$player\_down\_7_{it}$	3.388	8.161	0	171
$total\_favorites_{it}$	15.866	29.882	0	184
$leader\_up\_7_{ijt}$	0.016	0.134	0	13
$leader\_down\_7_{ijt}$	0.016	0.145	0	23
$follower\_up\_7_{ijt}$	0.010	0.059	0	10
$follower\_down\_7_{ijt}$	0.017	0.144	0	19
$leader\_positive_{ijt}$	0.003	0.059	0	10
$leader\_negative_{ijt}$	0.003	0.061	0	13
$follower\_positive_{ijt}$	0.002	0.048	0	7
$follower\_negative_{ijt}$	0.004	0.062	0	9

For all variables, sample size $n = 85,079,589$.
Variable subscripts describe their level of variation in the data. Trading days are indexed by t, players are indexed by i, and securities are indexed by j. All variables capture information available at the beginning of trading day t.

.3 CAPS Expanded Network Plot

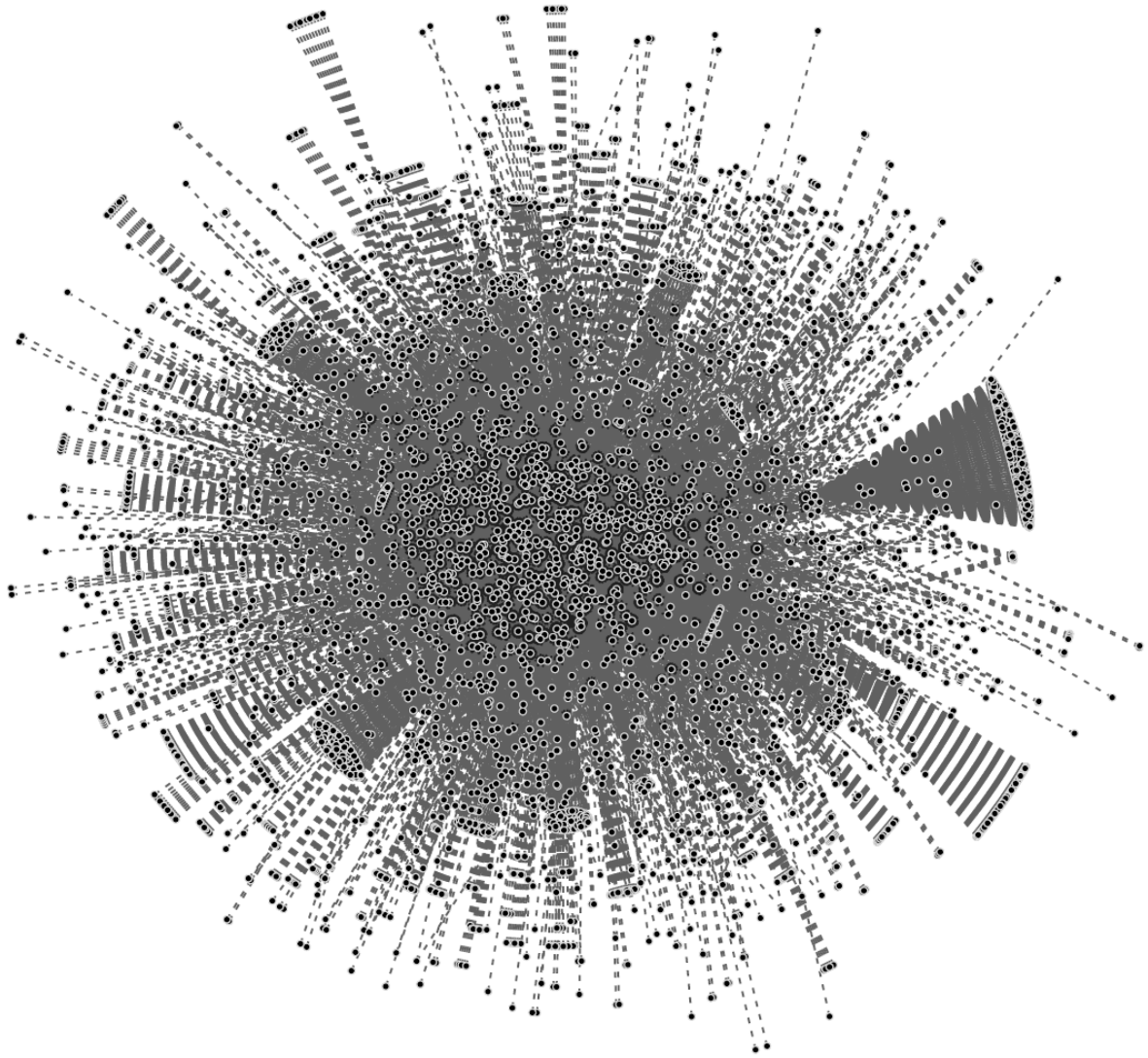

Figure 5: Extended Social Network - December 2013

.4 Steam Variable Definitions

Notation	Variable	Definition
q_{jt}	Aggregate Demand	The aggregate quantity demanded for game k on day t.
q_{ijt}	Individual Demand	The demand $\in \{0, 1\}$ for individual i of game j.
p_{jt}	Price	The price of game j on day t.
X_{jt}	Ownership Share	The proportion of active customers who own game j.
	Game Age	The number of months since a game's initial release.
	Median Hours Played	The number of hours the median owner has spent playing game j.
	Supported Languages	The number of officially supported game languages.
	Single-Player	Game provides a single-player game-play experience.
	Multi-Player	Game provides a multi-player game-play experience.
	Cooperative	Game provides a cooperative game-play experience.
	Massively Multi-Player	Game provides a massively multi-player game-play experience.
	Virtual Reality	Game has virtual reality headset support.
	Trading Cards	Game includes Steam trading cards which may be redeemed for badges.
	Tracked Stats	Game includes competitive statistics tracking and leader-boards.
	Mod Support	Game supports the use of user-submitted modifications.
	Has Achievements	Game features unlockable Steam achievements.
S_{jt}	Recent Positive Reviews	The number of favorable user-submitted reviews in the past week.
	Recent Negative Reviews	The number of unfavorable user-submitted reviews in the past week.
	Recent Sentiment	The proportion of recent reviews which are favorable.
	Cumulative Sentiment	The proportional of total reviews which are favorable.
	Twitch Channels	The number of broadcast channels for game j in the past week.
	Twitch Viewers	The number of viewers in thousands watching channels for game j in the past week.
	YouTube Videos	The number of uploaded YouTube videos for game j in the past week.
	YouTube Views	The number of views in thousands for uploaded YouTube videos in the past week.

Table continued on the following page.

Table 5: Variable Definitions - Continued

Notation	Variable	Definition
Z_{jt}	Day of Week	A day-of-week seasonal effect. (7 total)
	Month	A month-of-year seasonal effect. (12 total)
	Genre	The game's designated genre. (32 total)
	Tag	Extracted text tags from user-provided reviews. (337 total)
	ESRB Rating	The suitable audience rating provided by the Entertainment Software Rating Board. (6 total)
	Publisher	The publishing company with rights to distribute a game. (5,036 total)
	Developer	The development studio who designed and created the game. (6,681 total)
S_{ijt}	Degree	The number of current neighbors in a user's friendship network.
	Degree^2	The number of friends-of-friends in a user's friendship network.
	Support	The proportion of edges in a user's network which are supported by at least one mutual friend.
	Peer Purchases	The number of friends who have purchased a game in the past 7 days.
	Peer Ownership	The number of friends who already own a game.
	FoF Purchases	The number of 2^{nd} degree neighbors who have purchased a game in the past 7 days.
	FoF Ownership	The number of 2^{nd} degree neighbors who already own a game.
	Supported Purchases	The number of supported friends who have purchased a game in the past 7 days.
	Supported Ownership	The number of supported neighbors who already own a game.

.5 *Steam Aggregate Summary Statistics*

Table 6: Aggregate Estimation Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.
Aggregate Demand	415.615	30465.384	0	8799538
Price	10.442	10.831	0.99	399
Ownership Share	1.634	5.331	0.001	100
Game Age	14.897	10.557	0.033	51.467
Median Hours Played	3.617	11.504	0	2442.45
Supported Languages	3.603	4.054	1	27
Single-Player	0.963	0.188	0	1
Multi-Player	0.078	0.267	0	1
Cooperative	0.097	0.296	0	1
Massively Multi-Player	0.009	0.094	0	1
Virtual Reality	0.02	0.14	0	1
Trading Cards	0.59	0.492	0	1
Tracked Stats	0.884	0.861	0	3
Mod Support	0.061	0.239	0	1
Has Achievements	0.623	0.485	0	1
Recent Positive Reviews	9.845	133.151	0	56203
Recent Negative Reviews	2.487	64.532	0	45038
Recent Sentiment	60.85	28.273	0	100
Cumulative Sentiment	69.065	22.918	0	100
Twitch Viewers	0.058	1.191	0	257.257
Twitch Channels	1.863	32.643	0	9085
YouTube Views	0.594	8.698	0	1404.679
YouTube Videos	2.825	9.537	0	50

Statistics presented for 3,159,894 observations across 11,604 games.
Additional dummy variable features are not shown in the interest of brevity.

.5.1 Features Considered for CAPS Sentiment Weight

Variable	Definition
$success_{ijt}$	Dependent variable, successful prediction
$excess_{jt}$	Prior excess return generated by the prediction
$cumulative\_excess_{jt}$	Cumulative excess returns generated by this prediction to-date
$alpha_{jt}$	Estimated CAPM 3-year weekly alpha
$beta_{jt}$	Estimated CAPM 3-year weekly beta
$volume_{jt}$	Weekly trading volume
$shares_{jt}$	Total shares outstanding
$earnings_{jt}$	Prior quarter earnings per share
$gains_{jt}$	Prior quarter capital gains
$dividend_{jt}$	Expected dividend per share
ipo_{jt}	Prior year IPO dummy
$sector_j$	Fixed effects for 11 market sectors
$exchange_j$	Fixed effects for 7 traded exchanges
$picks_{it}$	Number of total predictions to-date made by player
$success\_rate_{it}$	Historical prediction success rate for player
$sector\_emphasis_{ijt}$	Share of historical predictions in market sector
$sector\_success_{ijt}$	Historical sector-specific success rate
$exchange\_emphasis_{ijt}$	Share of historical predictions on traded exchange
$exchange\_success_{ijt}$	Historical exchange-specific success rate
$tenure_{it}$	Platform tenure in weeks
$favorites_{it}$	Number of active social connections
$player_i$	Fixed effects for 946 players
age_{ijt}	Age of open player position with asset j
age^2_{ijt}	Squared-age of open position
$pitch_{ijt}$	Whether a prediction was accompanied with a text-based pitch
$horizon_{ijt}$	Indicator for whether recommended horizon of the prediction

Variable subscripts describe their level of variation in the data. Weeks are indexed by t, players are indexed by i, and securities are indexed by j.

.6 CAPS Sentiment Weight Coefficient Estimates

Covariate	Coef.	Std. Err.	z	P-value
lr	-0.0778719	0.01033	-7.54	0
lR	0.3950303	0.0248397	15.9	0
alpha	-2.425852	0.0233283	-103.99	0
beta	-0.003701	0.0006504	-5.69	0
div	-0.0037514	0.0181047	-0.21	0.836
shrout	-3.19E-13	9.75E-14	-3.27	0.001
yh_volume	3.67E-10	2.79E-11	13.15	0
eps	-0.0000204	0.0000215	-0.95	0.343
capgn	0.0078909	0.003651	2.16	0.031
recentipo	0.2203217	0.0064513	34.15	0
npicks	-0.000021	2.33E-06	-9.01	0
success_rate	0.0729458	0.0307088	2.38	0.018
nsector	0.0000464	5.56E-06	8.34	0
sector_share	-0.074554	0.0097317	-7.66	0
sector_rate	0.4053148	0.0155387	26.08	0
nexchange	0.0000356	7.59E-06	4.7	0
exchange_share	-0.0299923	0.0098328	-3.05	0.002
exchange_rate	0.0282304	0.0136416	2.07	0.039
tenure	0.0000142	9.80E-06	1.45	0.147
total_favorites	0.1593587	0.019673	8.1	0
constant	-1.673875	0.176461	-9.49	0
Player Fixed Effects	yes			
Sector Fixed Effects	yes			
Exchange Fixed Effects	yes			

Variable subscripts describe their level of variation in the data. Weeks are indexed by t, players are indexed by i, and securities are indexed by j.

VITA

Andrew Clayton is a graduate of the University of Washington, now working as an economist at Amazon.com. Any comments are welcome to aaclayton@gmail.com. Further details available at https://www.linkedin.com/in/andrew-clayton/.