ESSAYS ON CONSUMER PREFERENCES ONLINE

by

Maxwell C. Deaton

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Economics

Middle Tennessee State University
May 2018

Dissertation Committee:

Dr. Adam Rennhoff, Chair
Dr. Michael Roach
Dr. Anthon Eff

To Erin.

ACKNOWLEDGEMENTS

I owe a debt of gratitude to many people who helped me complete my doctoral studies. First and foremost I would like to thank my wife, Erin, whose love and encouragement have made this endeavor possible. I am also grateful for the endless support of my family. Without you I would not be where I am today. Finally, I would like to thank the excellent professors whose training and advice were invaluable. Thank you Charles Baum, Jason DeBacker, Stuart Fowler, Karen Mulligan, Mark Owens, and Joachim Zietz. I am particularly thankful to my committee for their patience, assistance, and advice. Thank you Adam Rennhoff, Mike Roach, and Tony Eff.

ABSTRACT

This dissertation is composed of three separate empirical analyses. Each is a separate analysis and article.

Theory suggests information about quality is especially important for experience goods. However, in an environment with multiple sources of information it is unclear which sources of information will be valued and if those sources will be valued differently conditional on product characteristics. Chapter I examines the impact of heterogeneous sources of information about quality on the demand for experience goods. In particular, I study the effects of user and expert reviews on the demand for video games online and how the valuation of information may vary based on if the game is produced by an 'independent' or 'major' developer. Using a unique dataset from the online video game platform Steam, I find consumers do value information about quality, they value information about quality from both user and expert reviewers, and they value user and expert reviews for products from independent developers and for products from major developers.

As individuals consume an increasing amount of news on social media it is important to understand consumer preferences, a key component of demand, for individuals on this platform. Chapter II examines the importance of ideological similarity between a news source and an individual and the importance of the reliability of a news source in individual's consumption decisions in the news market on social media. The findings suggest, conditional on who an individual chooses to follow, liberal individuals have an increasing preference for news as it becomes more conservative and no preference for news as it becomes more liberal. Liberals are also found to have a preference for less reliable news which may be a preference for

sensationalism. Alternatively, conservatives have no preference for more liberal or more conservative news and have no preference for more reliable or less reliable news.

Chapter III provides greater insight into Twitter accounts, account networks, and activity on the platform. I find, on average, active US Twitter accounts have an audience of 6,000 accounts and receive information from 1,500 accounts. High-profile accounts have much larger audiences than lower-profile accounts. High-profile accounts constitute a very low percentage of the total number of accounts. Both high and lower-profile accounts have similar levels of activity on Twitter. Activity related to news constitutes a small percentage of total activity on Twitter. Finally, the majority of activity on Twitter is reacting to already posted content, and verified accounts are more likely to post original content than non-verified accounts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

HETEROGENEOUS INFORMATION ABOUT QUALITY AND THE DEMAND FOR EXPERIENCE GOODS: A STUDY USING THE ONLINE VIDEO GAME PLATFORM STEAM

1 Introduction

Information is a powerful resource. With the increasing prevalence of the internet, the cost of obtaining information is decreasing. This has the potential to improve the utility of consumers, particularly in experience goods markets where little is known about a product prior to consumption. Economic theory suggests obtaining information about the quality of an experience good before deciding whether to consume the good allows an individual to increase their expected utility (Nelson 1970). Simply put, if the cost of information decreases then utility should increase. This suggests information about quality should be an important component in the demand for experience goods.

A potentially large benefit of the internet is the heterogeneous sources of information. These sources may provide individuals with varying viewpoints to be used in consumption decisions. Additionally, with a large number of sources of information, consumers may value some sources of information and not others conditional on product characteristics. Some sources may be seen as more reliable for different types of goods. Alternatively, some sources of information may be considered unreliable in general and may therefore be of little use to consumers.

The online video game market is one in which information may be of particular value. The nature of games is such that consumers are unlikely to know the utility a product will provide prior to experiencing the good. With thousands of products available, the task of choosing the good which gives the highest utility becomes increasingly difficult. In addition, there is a large degree of variation in the types of products available on the online video game market, potentially increasing the value of heterogeneous sources of information. Information about the quality of a game can, therefore, be especially useful to those who participate in this market.

In this study, I examine the Steam online video game platform. There are two sources of information, user reviews and expert reviews, readily available to those participating in the market. Games on Steam are produced either by “independent” or “major” developers. An independent developer is one which is not financed by a publisher (Rosen 2009). This suggests independent developers have fewer funds at their disposal than major developers. I expect games from major developers are likely to have better graphics, larger virtual worlds, and more choices and actions available to the player. Games from independent developers should be simpler. Another possibility is independent developers may have more creative freedom as a result of a lack of corporate oversight. As a result, games produced by independent developers may be particularly unique and creative in order to differentiate themselves from other products on the market. For these reasons, I expect games from independent and major developers will differ fundamentally in their characteristics.

The question I attempt to answer is, in an experience goods market with heterogeneous sources of information, what sources of information about quality do consumers value and does this valuation vary dependent on product characteristics? The fundamental differences in the characteristics of products from independent and major developers combined with the heterogeneous sources of information available on the Steam market provides an opportunity to gain insight into this issue. Specifically, I attempt to answer three questions: 1) Do user or expert reviews have an impact on the demand for products in the Steam market? The results suggest user and expert reviews are valued by consumers. 2) If so, do consumers value both types of information? Findings from the model imply both sources of information are valued. 3) Do consumers value different sources of information conditional on the type of product? In other words, do consumers value user reviews for certain products and expert reviews for other types of products (e.g. do consumers value user reviews for games from independent developers and expert reviews for games from major developers)? The results suggest the answer is no. Both user and expert reviews are valued on certain levels for products from independent developers and for products from major developers.

There is a long list of work which has studied the impact of one source of information about quality on demand for experience goods.[1] An exception to only considering one source of information is the work of Zhang and Dellarocas (2006). They look at the effect of professional (expert) and amateur (user) reviews on demand in the movie industry. They are unable to conclusively say if professional reviews have an effect on demand. They run a separate model to examine the effect of user reviews and find user reviews have a significant influence on the demand for movies.

Two other related studies look at user reviews, sometimes referred to as consumer reviews, and how they vary conditional on product characteristics. Zhu and Zhang (2010) look at

[1] See Eliashberg and Shugan (1997), Reinstein and Snyder (2005), Chevalier and Mayzlin (2006), and Shao (2012).

how online consumer reviews have different effects on demand based on product and consumer characteristics in the console video game market. They find online consumer reviews have a larger effect for products which are not as popular and for products whose users have more internet experience. Another related study, Blal and Sturman (2014) find user scores have a positive impact on the demand for luxury hotels but not for lower end products. Both of these studies suggest reviews may have a varying effect dependent on the product characteristics. The current study is the first, to my knowledge, to examine how alternative sources of information may be valued differently based on product characteristics.

The remainder of this paper is organized as follows. Section 2 lays out the theoretical framework. Section 3 describes the experience goods market on Steam. Section 4 explains the data collection process. Section 5 introduces the empirical strategy. Section 6 gives the results and discussion. Section 7 concludes.

2 Theoretical Framework

Video games, the focus of this study, can be defined as experience goods. An experience good is one for which it is difficult for a consumer to know their utility from consuming the good prior to consumption. Nelson (1970) has established a basic theory of experience goods. According to his theory, consumers are willing to purchase information about a good by experience as long as the marginal benefit of the information exceeds the marginal cost. Information is purchased by consuming, or experiencing, a good in an experience goods market. If it is assumed the consumer has no information about an unexperienced product, the choice of which product to experience becomes random. The marginal cost of experimenting with a random and unknown product is the difference between the expected utility of the best known option, an already experienced good for which the consumer has information, and the expected utility from continued random sampling.[2] Consumers can improve their expected utility by using "guided sampling."[3] In this situation individuals seek information from sources which have already experienced a good to inform their own decision about which good to experience. By increasing the expected utility of experiencing a given good, the marginal cost of experiencing the good is reduced.

There can be a great deal of heterogeneity in the value and availability of information. The internet has made the cost of finding information cheaper and, as a result, is a potentially important source of information, particularly information concerning the quality of experience goods. The internet has also reduced the cost of providing information potentially causing the quality of information available to vary widely. When seeking information for purchasing decisions, consumers have the option to rely on a multitude of sources, an individual source, or none at all. Some sources may be perceived as more reliable than others (e.g. expert reviewers could be seen as more knowledgeable sources of information). Alternatively, different sources of information could be considered more reliable than others conditional on additional information, such as product characteristics (e.g. user reviews are valued for games from independent developers and expert reviews are valued for games from major developers). For example, the reviews of users may be thought of as more reliable for little known, independently developed games because users may seek out and play relatively obscure games and have tastes which more accurately reflect the preferences of potential consumers of this type of product than expert reviewers. On the other hand, expert reviewers may be seen as more reliable in this environment

[2] The expected utility of continued random sampling is the mean of the utility distribution.
[3] Assuming positive correlation in tastes, the expected utility of a recommended sample is higher than the expected utility of a random one.

because of their perceived above average knowledge of the industry. Additionally, consumers may value information more for certain products than others. Because independently produced games likely vary more in quality than those produced by larger developers, all sources of information may be valued more for consumers deciding whether to experience goods from independent developers. Alternatively, all reviews may be valued only for games from major developers because these games may be seen as having higher quality and thus providing higher utility than games from independent developers. Here consumers may narrow their search to major developers first and use reviews to select the product from this smaller sample which provides the highest utility. Finally, consumers may decide that the information provided on the internet is not reliable enough to factor into decisions about which goods to experience.

Because information about the quality of experience goods potentially impacts the expected utility of consuming a new experience good, it is an important component of demand. If a consumer expects information to be reliable, then a positive review should have a positive effect on demand and vice versa.

3 Steam Market

Steam is an online entertainment platform, founded and run by Valve Corporation since 2003 (Plunkett 2013), which is primarily used for purchasing and playing PC video games. In order to purchase and download games from Steam, it is necessary to create a profile on the platform. The profile stores information on the games owned and user statistics, such as the amount of time a game has been played.

Steam is well suited for studying the impact of information about quality on demand for several reasons. First, it has a large number of users, 125 million in 2015 (Saed 2015), and a large number of products available, over 4000 for PC (Larabel 2015), providing a large number of observations. It is one of the most popular gaming platforms in the world implying, in a very real sense, it is *the* market for PC games. Perhaps the most appealing aspect of Steam, however, is the prominence of user and expert reviews available to the consumer. When a consumer purchases a game on Steam they must go to the game's page. Both user scores and Metacritic (expert) scores are displayed on a games page. User scores are displayed in categories such as “negative” or “overwhelmingly positive” so while the consumer does not immediately see the percentage of positive reviews, they do immediately receive information about how consumer reviewers generally view the quality of a game. Steam also posts the Metacritic score, if available for a game, directly to the game's page. The Metacritic score is not displayed as prominently as the user score. Figures 1.1 and 1.2 give an example of a game's page and how the user and Metacritic scores are presented.

The fact information about quality is displayed so prominently on the Steam market gives the current study a distinct advantage over previous ones. For movies, console video games, etc. reviews must be sought out. While the cost of this search is relatively low, there is still the possibility a significant portion of consumers are not receiving information about quality. On the Steam market, consumers have this information broadcast to them before they purchase a given product. This implies I can be certain individuals obtain information about quality before making a consumption decision and strengthens the reliability of any causal impact I find.

It should be mentioned games have reduced prices fairly frequently on Steam. This includes events such as the Steam Summer Sale which is an annual sale during which thousands of products are offered at a reduced price. Because the demand model controls for price, however, I do not expect this to be a large concern. Below I discuss in detail how the data is constructed to account for any discount effects.[4]

4 Data

Data was gathered weekly from May 19, 2016 until November 17, 2016 giving 26 weeks of data. To construct a week of data, two points in time are needed. For example, the first week of data is generated using information on May 19 and May 26. This is necessary because information on quantity sold during a week is not given directly. Instead, data on total owners at a given time is collected and used to calculate weekly quantity. The exact methods I use to construct the data are discussed below.

Data such as price, metacritic score, genre tags, release date, product type, and the minimum PC memory requirement of products come directly from Steam via its application program interface (API) (Steam Community 2016). Not all games have price data, including free games. If a game has no price data and is either tagged as in the ‘Free to Play’ genre or Steam has tagged the game as free, I give the game a price of zero. This same information is used to create an indicator variable for free games. I remove entries without any price information.[5]

---

[4] For a better understanding of Steam please visit Steam (2018).

[5] These are removed at the same time I drop games without information about PC memory. This drop results in a loss of 5.4% of quantity sold and thus I do not expect this to be a significant issue.

Because the price can vary from the beginning of the week to the end of the week (e.g. during the Steam Summer Sale) and exact information on when the price changes is unobserved, I assume the price of a good for a given week is the lower price to capture the effect of sales.

Metacritic is a website which gathers and aggregates information on published reviews assigning games a metascore (Metacritic 2016). I assume the metascore to be an accurate representation of the general consensus of expert reviewers and thus metascores are used as the measure of expert information about quality. Metascores are given on a scale of 1 to 100 with higher scores representing better reviews. Because many games do not have a metascore, dummy variables are created indicating if a game has no metacritic score, a low metacritic score (below 50), or a high metacritic score (75 to 100) with middling scores excluded. This is beneficial for two reasons. First, it allows games with no metascore to be included in the analysis. Second, it helps determine if simply having a metascore is a signal which impacts consumer demand. In other words, if the availability of expert information is an important factor in demand.

Different developers have varying resources for promoting and developing their products. Some developers may have reputations signaling quality to consumers. Ideally, an indicator variable would be included for each developer, however, there are thousands of developers selling products on Steam and so it becomes unreasonable to control for them all. Instead, developers are divided into independent and major categories based on Steam's classification of the developer.[6] This is done under the assumption that major developers likely have greater resources to devote to creating and advertising a game as well as being more likely to have a widely established reputation.

[6] Games from independent developers are given the genre tag of “Indie” by Steam.

Information on the release date of a game is gathered and recoded as the number of months since release. This is done to control for the expected inverse relationship between the length of time a product has been available and weekly quantity sold. I define the market studied as the Steam market from 2010 onwards because older games are unlikely to have significant amounts of quantity sold recently.[7] Additionally, I drop games which have not been released and games released less than 3 days prior to the beginning of the week.[8]

Information concerning the minimum PC memory requirements of a game is also gathered. Games with larger environments, more detail, and a greater variety of actions and choices available to the player are assumed to be of higher quality on average. These games require more memory and so it is assumed PC memory is a good proxy for the quality of a game. The data is constructed in megabytes.[9] I drop games without information about PC memory.

Additional data on the number of owners, the score rank, the title, and a list of descriptive user tags of games is available using SteamSpy's API (SteamSpy API 2016). SteamSpy is a Steam stats service intended to supplement the information publicly available from Steam. To estimate the number of owners of a game on Steam, SteamSpy uses a polling technique where a number of user profiles on Steam are sampled at random daily and data is collected on the games which are owned. Polling is used because of the impractical proposition of daily collecting data

[7] I do not have data which goes back far enough to test this myself, however, support for this decision comes from Zhu and Zhang's (2010) study of the console video game market. They find the number of units sold drop drastically after the first few months. Although they study the console video game market, it seems likely a similar pattern would appear in the PC video game market.

[8] This is done because the estimates for owners of a game are conditional on 3 days of data (SteamSpy 2018). I conclude games with fewer than 3 days of data are unreliable. I drop games prior to 2010, games not yet released, and games released fewer than 3 days prior to the beginning of the week at the same time.

[9] Several games were recorded as having a memory size of 512 gigabytes. As this is an unreasonable size for a game, I recode these values as 512 megabytes.

from the millions of subscribers on Steam.[10] The technique has been checked against real-world data and has been shown to have reasonably accurate estimates (Orland 2014). It must be kept in mind, however, the data on owners is an estimate. Games which have an estimated number of owners of less than 30,000 are removed because these estimates have too few observations to be considered reliable (SteamSpy 2018).[11] Weekly quantity sold of a game is calculated as the number of owners at the end of the week minus the number of owners at the beginning of the week.[12] Because of the sampling process used to calculate the number of owners, if the quantity sold during a week is extremely low or zero, the number of owners estimated may be lower at the end of the week than at the beginning. This implies the quantity for a game may be calculated as a negative value. For entries with a negative value it is assumed quantity sold during the week is zero and therefore these quantities are recoded as zero. I drop the week between October 11 and October 20, 2016 from the study as during the middle of this week SteamSpy underwent a recalibration which impacted the calculated quantity sold.[13]

SteamSpy also gathers data called the score rank of a game. The score rank is defined as the percentage of games which have a lower Steam user score than the given game. For example, if a game has a score rank of 80, the game has a user score higher than 80% of the games on Steam. Steam user scores are numbers 1-100 indicating the percentage of positive user reviews. Individuals reviewing a game on Steam must select whether they recommend or do not

---

[10] The technique used by SteamSpy was developed by individuals at Ars Technica. For an in depth description of the process see Orland (2014).

[11] This is the first drop performed in cleaning the data. During the first week, the drop results in a decrease of 2.2% of the total owners in the market. This suggests the data dropped will not cause any significant issues.

[12] This requires dropping data not present in both weeks. I expect the products dropped are either hovering around 30,000 owners, and therefore are unreliable, or are new releases.

[13] See this Tweet from Sergey Galyonkin creator of SteamSpy: https://twitter.com/Steam_Spy/status/786944728604418050 (Galyonkin 2016).

recommend the game, giving the positive or negative user reviews. Reviewers also have a space to provide an explanation for why they do or do not recommend a game. To review a game on Steam an individual must have the game in their library. This is intended to ensure reviews are written by individuals who have played the game and intended to increase trust in the user score. During the middle of my data collection, Steam removed reviews from users who obtained a copy of the game through a Steam key, an access key occasionally handed out by developers, from impacting user scores (Yin-Poole 2016). This change is intended to keep developers from artificially boosting user review scores. This change occurred between September 15 and September 22, 2016 in my data. The Pearson correlation coefficient between the score ranks for games on September 15 and on September 22 which both have a user score is 0.997. Additionally, only two games which have a score rank in on September 15 did not have one in September 22. These findings suggests the change will not dramatically affect the results.[14] Score rank is used as the measure of user generated information about quality. I drop games without a value for score rank.[15]

Any product on Steam labeled as being in a utility genre, including educational, art and design, and similar products, is removed from the analysis. Additionally, I drop VR games and TV episodes. These products are removed based on the assumption utility products, VR games, and TV episodes are differentiated enough from video games to be considered a separate market.[16]

---

[14] The games used for the Pearson correlation coefficient are before any data is dropped.
[15] This results in a loss of 0.9% of quantity sold.
[16] VR games and TV episodes are dropped using information on title, type, and user tags from SteamSpy API (2016). Utility products are dropped using information from Steam's API (Steam Community 2016). Additionally, I drop several DLC and movies which I find as well as any multiplayer versions of games which have both single and multiplayer versions to avoid double counting.

The data from these two sources is merged to create the final sample. When constructing the data for an individual week, except for price and quantity, the values used are those for the beginning of the week. This is particularly important for data on score rank and metascores. To be able to study if a user or expert review has an impact on the demand for a game during a given week, it is essential the review score be from prior to consuming the game. Only reviews read before the purchase of a game have the potential to impact a consumer's decision about whether to experience the good. The final sample includes 73,094 product-week observations over 25 weeks.

5 Empirical Methodology

The empirical methodology I use follows the discrete choice demand model developed in Berry (1994) and Berry, Levinsohn, and Pakes (1995) which is sometimes refered to as the BLP model of demand. Practical advantages of the model include allowing for estimation with aggregate data, heterogeneity in consumer preferences, estimation with endogenous prices, and more realistic substitution patterns than other demand models. Additionally, the model has the advantage of being rooted in the utility function of consumers, as suggested by economic theory, allowing for marginal utility estimates for product characteristics to be obtained. I present a brief overview of the model and estimation procedure which closely follows the descriptions of Nevo (2000) and Vincent (2015). I also discuss the instruments used in this study as well as the variables included in the mean utility function.

5.1 Model

Assume $t = 1, \ldots, T$ markets are observed with each containing $i = 1, \ldots, I$ consumers. In each market, product prices, characteristics, and aggregate quantities are observed for each of the $j = 1, \ldots, J_t$ products. The subscript t is used because product offerings are not identical across markets in my sample (Nevo 2000). In this study, a market is defined as the Steam market for a given week. Demand specification under the BLP model has three primary components.

First, consumer i's conditional indirect utility from consuming product j in market t which is a function of both observed and unobserved product and individual characteristics is specified. The utility can be written as:

$$u_{ijt} = \alpha_i(y_i - p_{jt}) + x_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \text{ , (1)}$$

where y_i is consumer i's income, p_{jt} is the price of good j in market t, α_i is a parameter which captures the marginal utility of income, x_{jt} is a K-length vector of product j's characteristics in market t, β_i is a K-length column vector of parameters which capture the marginal utility of the product characteristics, ξ_{jt} are the unobserved characteristics of product j in market t, and ε_{ijt} is a stochastic term with a mean of zero. Notice the parameters are specific to each consumer i. An assumption important in the estimation process is the product characteristics, x_{jt}'s, are mean independent of the unobserved product characteristics, ξ_{jt}.

The second component of the model uses characteristics of individuals to introduce variation in consumer preferences. This is done by modeling α_i and β_i as functions of demographics and unobserved individual characteristics:

$$\binom{\alpha_i}{\beta_i} = \binom{\alpha}{\beta} + \Pi D_i + \Sigma v_i \text{ , (2)}$$

where α and β are mean values of the given taste parameters, D_i is a $d \times 1$ vector containing information about observed demographics, Π is a $(K+1) \times d$ matrix of parameters which show how the taste for product characteristics vary with demographics, v_i captures individual characteristics unobserved by the econometrician, and Σ is a $(K+1) \times (K+1)$ matrix of parameters for v_i. It is assumed $D_i \sim \hat{P}_D(D)$ and $v_i \sim P_v(v)$. As both D_i and v_i are unobserved, $\hat{P}_D(D)$ and $P_v(v)$ are assumed to be parametric distributions.

The conditional indirect utility can be rewritten when the first two equations are combined giving:

$$u_{ijt} = \alpha_i y_i + \delta_{jt} + \mu_{ijt} + \varepsilon_{ijt} \text{ , (3)}$$

where $\delta_{jt} = x_{jt}\beta - \alpha p_{jt} + \xi_{jt}$ and $\mu_{ijt} = (-p_{jt},\ x_{jt})(\Pi D_i + \Sigma v_i)$, which is a $1 \times (K+1)$ vector. δ_{jt} is the mean utility of product j in market t. Notice there are no i subscripts implying this term does not vary over individuals. Heterogeneity enters the equation through μ_{ijt} and ε_{ijt} which vary over individuals. μ_{ijt} and ε_{ijt} have a mean of zero. They are heteroskedastic divergences from the mean utility for product j in market t.

The final component needed for the demand model is the outside good. The outside good is the option to not buy any product in the market. The conditional indirect utility of the outside option is given by:

$$u_{i0t} = \alpha_i y_i + \xi_{0t} + \pi_0 D_i + \sigma_0 v_{i0} + \varepsilon_{i0t} \text{ . (4)}$$

I follow the standard practice and assume ξ_{0t}, π_0, and σ_0 are all equal to zero. Because $\alpha_i y_i$ is common to all products in the market, including the outside option, the term will disappear from this function as well as all product utility functions. This implies the utility from the outside option is zero.

A necessary assumption is that all consumers purchase one unit of whichever good gives them the highest utility. This assumption implies consumers choose a product such that $u_{ijt} > u_{ilt}, \forall l = 0, \dots, J, l \neq j$. As long as this condition holds true, the set of characteristics which define the individuals who choose product j can be expressed as $A_{jt} = \{D_i, v_i, \varepsilon_{it}\}$ where $\varepsilon_{it} = \{\varepsilon_{i0t}, \dots, \varepsilon_{iJt}\}$. The market share of product j can be calculated by taking the integral over A_{jt}. Assuming ties do not occur, the predicted market share of product j in market t, denoted as s_{jt}, can be expressed as:

$$s_{jt} = \int_{A_{jt}} dP_{\varepsilon}(\varepsilon) dP_v(v) d\hat{P}_D(D) \text{ , (5)}$$

where $P()$ are the population distribution functions.

Assuming ε_{ijt} are distributed i.i.d. with Type 1 extreme-value, the probability of consumer i choosing product j in market t can be expressed as:

$$s_{ijt} = \frac{exp(\delta_{jt} + \mu_{ijt})}{1 + \sum_{k=1}^{K} exp(\delta_{kt} + \mu_{ikt})} \text{ . (6)}$$

Here, μ_{ijt} allows for correlation between choices because of the inclusion of demographics and individual characteristics. The correlation allows for more realistic substitution patterns between products.[17]

To get the marginal utility estimates from the model, (5) must be solved. However, this integral cannot be evaluated analytically. Instead, the integral is approximated by Monte Carlo integration with R random draws of D_i and v_i from their distributions.[18] The market share of product j can now be expressed as:

$$s_{jt} = \frac{1}{R} \sum_{i=1}^{R} s_{ijt} \text{ . (7)}$$

---

[17] See Nevo (2000) Sec. 2.2 for an in depth discussion of the distributional assumptions placed on ε_{ijt} and how the inclusion of μ_{ijt} allows for more realistic substitution patterns.
[18] This is discussed in some detail in Vincent (2015).

5.2 Estimation

What follows is a brief description of the estimation procedure as described by Vincent (2015).[19] To estimate the model a Generalized Method of Moments (GMM) estimation process is used. The GMM objective function is created by interacting the instruments for price and the random coefficients, z_{jt}, and the error term, ξ_{jt}, implied for given values of the unknown model parameters.[20] Defining z_{jt} to be a set of instruments which are functions of product characteristics gives us the moment restriction:

$$E(z_{jt}\xi_{jt}) = 0, \quad (8)$$

for all products. [21] This will allow identification of the model parameters. Below are the steps to solve for the GMM estimator as presented in Vincent (2015): First, to approximate the integral from (5) start by taking R random draws for the demographic variables, D, and for the unobserved individual characteristics, ν. Next, solve for the mean utility, δ_{jt}, for all products in each market such that the predicted market shares, s_{jt}, from (5) equal the observed market shares, S_{jt}, for given values of Π and Σ. This is done via contraction mapping. Third, find the GMM objective function by calculating the sample moment conditions $\frac{1}{T}\sum_{t=1}^{T} Z_t'\xi_t$. Z_t is a $J \times l$ set of instruments and ξ_t is the set of mean utility errors for all products in a given market, t.

---

[19] For an in depth explanation of the estimation procedure see Vincent (2015) Sec. 3 and Nevo (2000) Sec. 3.3.
[20] See Nevo (2000) Sec. 3.3.
[21] Vincent (2015) defines z as a function of product characteristics and cost-shifters. I do not have data for cost-shifters and so rely on product characteristics. This is discussed further in Sec. 5.3.

Finally, find the values of the parameters of both observed characteristics, α and β, and unobserved characteristics, Π and Σ, which minimize the GMM objective function.

The off-diagonal elements of the covariance to be zero. This implies any correlation between taste parameters are driven entirely by demographics (Vincent 2015).

5.3 Instruments

A key feature of the BLP model of demand is it allows the endogeneity of price to be addressed. The instruments allow both the coefficient for price and the variance of random parameters to be identified. My model only includes one random parameter on price.

The instruments I use include the square of all product characteristics, x_{jt}^2, and the sum of all characteristics of other products, $\sum_{m=1,m\neq \mathrm{j}}^{J_t} x_{mt}$. These are similar to the 'standard instruments' mentioned in Vincent (2015) and are common in the literature.

5.4 Mean Utility for Steam

Here I discuss the variables included in the mean utility function. To study the impact of varying sources of information about quality on the demand for video games I run the demand model using the following mean utility function for product j in market t:

$$\delta_{jt} = \beta_0 + ScoreRank_{jt}\beta_1 + ScoreRank_{jt}^2\beta_2 + NoMetascore_{jt}\beta_3 + LowMetascore_{jt}\beta_4 + HighMetascore_{jt}\beta_5 + Independent_{jt}\beta_6 + ScoreRank_{jt} * Independent_{jt}\beta_7 + ScoreRank_{jt}^2 * Independent_{jt}\beta_8 + NoMetascore_{jt} * Independent_{jt}\beta_9 + LowMetascore_{jt} * Independent_{jt}\beta_{10} + HighMetascore_{jt} * Independent_{jt}\beta_{11} + PCMemory_{jt}\beta_{12} + PCMemory_{jt}^2\beta_{13} + Months_{jt}\beta_{14} + Months_{jt}^2\beta_{15} + Free_{jt}\beta_{16} + \alpha Price_{jt} + \xi_{jt}. \quad (9)$$

As described above, $ScoreRank$ is the measure of user scores. The effect of not having an expert review is captured by the $NoMetascore$ dummy variable. The effect of a low or high expert review is captured by the $LowMetascore$ and $HighMetascore$ dummy variables respectively. Because estimating the effect of information about quality (reviews) on the demand for games, separate from inherent quality, is the goal of this study, it is necessary to control for quality. It is difficult to observe the quality of experience goods. As discussed in Section 4, I believe memory requirements for a game are a good proxy for quality. Therefore, $PCMemory$ is intended to control for the inherent quality of a game allowing the effects of reviews to be estimated separately from the effect of quality. $Months$ controls for the length of time a product has been available. $Independent$ is a dummy variable equal to one if the developer of the product was an independent developer. I include a dummy variable $Free$ to indicate whether a game is free or not. I suspect free games may signal low quality and thus have a negative impact on demand. However, free games may also be seen as having a low opportunity cost of experiencing the product and may have a positive impact on demand. $Price$ is the only random variable in the model. It is expected $ScoreRank$, $PCMemory$, and $Months$ enter the model non-linearly, thus squared terms are included in the mean utility function.

The mean utility function also has five interaction terms. The interaction terms are included to capture the differing effects of reviews on major and independent developers. However, the specification of the model requires some coefficients to be constructed manually. For example, while the coefficient on $LowMetascore, \beta_4$, can be interpreted as the effect of low expert reviews on major developers, the coefficient on $LowMetascore * Independent, \beta_{10}$, has a less straightforward interpretation. β_{10} is interpreted as the difference between the effect of all low expert reviews and the low expert reviews for games from independent developers. To find the effect of low expert reviews on games from independent developers, I sum β_4 and β_{10} to create a new coefficient. New coefficients and the associated standard errors are calculated for all the interactions between reviews and independent developers. The effects of reviews on independent games and the effects of reviews on major games are presented.

To run the model, it is necessary to have the observed market shares for each product in each market and the share of the outside good in each market. The market share of good j in market t is calculated as the quantity of good j in market t divided by the total size of market t, M_t. The market size is the total number of potential consumers in a market. The observed share of product j in market t can be expressed as $S_{jt} = \frac{quantity_{jt}}{M_t}$ with the share of the outside good given by $S_{0t} = 1 - \sum_{j=1}^{J} \frac{S_{jt}}{M_t}$. To calculate the observed market shares, an assumption must be made about the market size. There is a feature of the Steam market which can assist in making this assumption. As mentioned previously, in 2015 it was reported that Steam had over 125 million active users (Saed 2015). From this piece of information, I assume roughly a fourth of users are active in a given week. This gives me a total market size of 31.25 million consumers. I assume the size of the market is constant throughout all markets. There is another issue which must be addressed when calculating the market share of a product on Steam. In a market with thousands of available products, it becomes likely some of these products will have zero quantity in certain weeks. This is the case in my data. Because the model requires market shares be between zero and one, I use a shift to ensure all products have a non-zero market share. This is accomplished by adding a value of one to the quantity of each game, shifting the quantity of all products in the market.

6 Results and Discussion

Table 1.1 shows the descriptive statistics for user and expert reviews on Steam. There are a few things to note. First, if no data had been dropped, the mean of the score rank variable would be 50 by definition. The data in this study has a slightly lower average of user scores. However, the mean of the score rank variable is close enough to 50 to suggest the final sample is representative of the market at large. One striking feature of the market is 59.9% of games on Steam do not have an expert review. This is not entirely unexpected, however, because the large number of products on Steam make it implausible for all games to be reviewed by a limited number of publications. The extraordinarily small percentage of games, 1.3%, which have a low expert review is surprising. Games on Steam which have an expert review rarely receive a poor one. On the other hand, 18.8% of games receive a high metascore and 20.0% of games receive a middling metascore. This implies about half of games reviewed by an expert are the recipients of high praise, which suggests expert reviewers are prone to give positive scores. This may imply experts select games they believe will be of high quality to review and gives justification for the inclusion of the No Metascore variable.

Table 1.2 gives further descriptive statistics for products on Steam. Games average $9.84 per unit, however, there is a wide range of prices with the most expensive games running $59.99 and the cheapest being free. Over half of the games on Steam are produced by independent developers. Independent developers are often unknown and thus may not give a signal about the quality of their products. The potentially unknown nature of independent goods combined with the fact there are a high number of these products makes it difficult for consumers to discern which are of high quality. This suggests information about quality may be especially useful to consumers deciding whether to experience a product created by an independent developer. The average game requires 1.7 gigabytes of memory, however, games can have memories as large as 16 gigabytes and as small as 1/1000 of a gigabyte. Free games make up 1.59% of the total games on Steam and the average game has been out for slightly more than 2 and a third years.

Table 1.3 gives descriptive statistics for reviews of products from independent developers and Table 1.4 gives descriptive statistics for reviews of products from major developers. The average score rank for products from independent developers is higher than the average score rank for products from major developers. This suggests users review products from independent developers more favorably or, alternatively, review products from major developers more harshly. 64.9% of games from independent developers do not have an expert review while 53.1% of products from major developers have no expert review. In other words, expert reviewers review a higher portion of games from major developers than from independent developers. Products from both independent and major developers rarely receive a low expert review score. 16.3% of games from independent developers obtain a high metascore compared to 22.1% of games from major developers. This suggests games from major developers are reviewed more favorably by experts.

The results for the model are presented in two Tables. Table 1.5 presents the marginal utilities not related to reviews. Table 1.6 presents the marginal utilities for reviews by major and independent developers. It is important to keep in mind the coefficients presented are marginal utilities. Not all results presented. The results not reported are all of the expected sign.[22]

[22] The effect of PC memory is positive, significant, and decreasing. It is relatively small in magnitude. In other words, quality has a positive impact on demand, however, as quality increases the effect is diminished. The effect of months is negative, significant, and increasing. This implies time since release has a negative impact on demand, however, as more time passes the negative effect lessens.

The effect of a game coming from an independent developer is negative and significant relative to the effect of a game coming from a major developer. There are several possible reasons for this result. First, major developers are likely to have more resources to promote awareness of their products (i.e. advertising) and so it is possible these games have higher demand simply because of awareness. The second possibility is major developers have developed reputations for delivering products with a certain level of quality. In this scenario, a product from a major developer may be seen as less of a risk and so the cost of experiencing the good is lower. It could also be a combination of these options. The coefficient on free games is negative and significant. This suggests free games may signal low quality which offsets any reduced cost of experiencing the good. Finally, it should be mentioned the coefficient of price is negative and significant. This matches what is predicted by economic theory.

User scores have a positive and significant effect on demand. In other words, more positive user scores should result in higher demand. This holds true for both games from independent developers and games from major developers. The effect appears to be linear for games from independent developers and decreasing for games from major developers. It should be kept in mind the non-linear effect for major developers is small in magnitude. Not having an expert review has a negative and significant impact on demand for products from both independent and major developers This could be because consumers believe experts select high quality games to review and, therefore, having a metascore is a signal of quality. Alternatively, it could be only well known games are reviewed by experts, and games which do not have a metascore are simply unknown to consumers and therefore only consumed by a small section of the market. Interestingly, having a low metascore, for both independent and major developers, is not statistically different from zero suggesting low metascores are not different than middling metascores. A high expert review has a positive and statistically significant effect on the demand for a game relative to middling reviews for both types of developers. These results suggest consumers value information of different types and they value it similarly for games from both independent and major developers.[23]

The results of the demand model allow me to answer the three questions specified earlier. The answer to the first question, concerning whether consumers value user or expert reviews, is certainly yes. User reviews have a positive and significant effect on demand for products from both independent and major developers. Not having an expert review negatively impacts demand while having a high metascore has a positive effect on demand. The evidence supports the notion consumers value information.

The answer to the second question, addressing whether consumers value multiple sources of information, appears to be yes. While the study does not address the difference in magnitude of the valuations of user and expert reviews, consumers value both sources of information in their decision making process.

Finally, the results suggest the answer to the third question, concerning whether the valuations of heterogeneous sources of information change conditional on product type, is no. The results of the model are almost identical in terms of sign and significance. While the magnitudes differ in certain cases, the results do not suggest users reviews are only valued for one type of product and expert reviews are only valued for another. It should be kept in mind the evidence from the model is limited, however, because I only examine how the valuations of these heterogeneous sources of information vary on one level. It may be that consumers do vary

[23] It should be kept in mind the magnitudes of the marginal utilities, particularly for score rank and high metascores, are larger for games from major developers than for games from independent developers.

the source of information they value on other dimensions, such as genre. For example, consumers may value user reviews for games in genres with a high number of interactions with other players because of the strong network effects, while expert reviews may have a stronger effect for single player experiences. Therefore, while the current study suggests that consumers do not alter their valuations of sources of information by product type, this cannot be taken as a definitive answer.

7 Conclusion

The results of this study suggest information, online reviews in particular, is an important factor in the demand for experience goods. Consumers appear to value multiple sources of information. In particular, a higher score rank has a positive effect on demand. Not having an expert review negatively impact the demand for a good, while a high expert review score has a positive effect on demand. These results hold true for games from both independent and major developers.

This work highlights the importance of considering heterogeneous sources of information when examining how information about quality impacts demand. When there are varying sources of information available to consumers, they appear likely to value multiple sources. It is important for future research to keep this in mind when considering the impact of information about quality on the demand for experience goods.

REFERENCES

1. Berry, Steven. 1994. “Estimating Discrete-Choice Models of Product Differentiation.” RAND Journal of Economics 25(2):242-262.

2. Berry, Steven., James Levinsohn, and Ariel Pakes. 1995. “Automobile Prices in Market Equilibrium.” Econometrica 63(4):841-890.

3. Blal, Inès., and Michael Sturman. 2014. “The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales.” Cornell Hospitality Quarterly 55(4):365-375.

4. Chevalier, Judith., and Dina Mayzlin. 2006. “The Effect of Word of Mouth on Sales: Online Book Reviews.” Journal of Marketing Research 43(3):345-354.

5. Eliashberg, Jehoshua., and Steven Shugan. 1997. “Film Critics: Influencers or Predictors?” Journal of Marketing 61(2):68-78.

6. Galyonkin, Sergey. 2016. “I was informed by numerous devs that Steam Spy is roughly 5% off. It's time for another calibration. Brace for impact!” 14 October. Tweet. https://twitter.com/Steam_Spy/status/786944728604418050.

7. Larbel, Michael. 2015. “Steam Crosses 1,500 Games Natively Available for Linux.” Phoronix. https://www.phoronix.com/scan.php?page=article&item=steam-1500-linux&num=1.

8. Metacritic. http://www.metacritic.com/. Accessed 2016.

9. Nelson, Phillip. 1970. “Information and Consumer Behavior.” Journal of Political Economy 78(2):311-329.

10. Nevo, Aviv. 2000. “A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand.” Journal of Economics and Management Strategy 9(4):513-548.

11. Orland, Kyle. 2014. “Introducing Steam Gauge: Ars Reveals Steam's Most Popular Games.” Ars Techina. https://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/.

12. Plunket, Luke. 2013. “Steam is 10 Today. Remember When It Sucked?” Kotaku https://kotaku.com/steam-is-10-today-remember-when-it-sucked-1297594444.

13. Reinstein, David. and Christopher Snyder. 2005. “The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics.” The Journal of Industrial Economics 53(1):27-51.

14. Rosen, David. 2009. “What are Indie Games?” Wolfire Blog. http://blog.wolfire.com/2009/08/what-are-indie-games/.

15. Saed, Sherif. 2015. “Steam has over 125 million active users, 8.9M concurrent peak.” VG24/7. https://www.vg247.com/2015/02/24/steam-has-over-125-million-active-users-8-9m-concurrent-peak/.

16. Shao, KungHsin. 2012. “The Effects of Controversial Reviews on Product Sales Performance: The Mediating Role of the Volume of Word of Mouth.” International Journal of Marketing Studies 4(2):32-38.

17. Steam. http://store.steampowered.com/. Accessed 2018.

18. Steam Community. https://steamcommunity.com/dev. Accessed 2016.

19. SteamSpy. https://steamspy.com/. Accessed 2018.

20. SteamSpy API. https://steamspy.com/api.php. Accessed 2016.

21. Vincent, David. 2015. “The Berry-Levinsohn-Pakes Estimator of the Random-Coefficients Logit Demand Model.” The Stata Journal 15(3):854-880.

22. Yin-Poole, Wesley. 2016. “Valve finally takes on Steam user review score manipulation: But has it been heavy-handed?” Eurogamer. http://www.eurogamer.net/articles/2016-09-13-valve-takes-on-steam-user-review-score-manipulation.

23. Zhang, Xiaoquan., and Chrysanthos Dellarocas. 2006. “The Lord of the Ratings: Is a Movie's Fate Influenced by Reviews?” Twenty Seventh International Conference on Information Systems.

24. Zhu, Feng. and Xiaoquan Zhang. 2010. “Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics.” Journal of Marketing 74(2):133-148.

APPENDIX A: CHAPTER I TABLES

Table 1.1: Descriptive Statistics for Reviews

Variable	Mean	Std. Dev.	Min.	Max	Obs.
Score Rank	47.984	28.335	0	100	73,094
No Metascore	0.599	.490	0	1	73,094
Low Metascore	0.013	0.113	0	1	73,094
High Metascore	0.188	0.390	0	1	73,094

Table 1.2: Game Descriptive Statistics

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Price	9.835	10.217	0	59.99	73,094
Independent	0.579	0.494	0	1	73,094
PC Memory	1,663.989	1,330.651	1	16,000	73,094
Free	0.159	0.365	0	1	73,094
Months	28.848	19.153	0	82	73,094

Note: Price data is recorded in cents for use in estimation. I convert to dollars here for ease of interpretation. It should also be noted that PC Memory is reported in megabytes (1000 MB=1 GB).

Table 1.3: Descriptive Statistics for Reviews for Independent Developers

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Score Rank	50.762	28.003	0	100	42,300
No Metascore	0.649	0.478	0	1	42,300
Low Metascore	0.009	0.095	0	1	42,300
High Metascore	0.163	0.370	0	1	42,300

Table 1.4: Descriptive Statistics for Reviews for Major Developers

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Score Rank	44.168	28.344	0	100	30,794
No Metascore	0.531	0.499	0	1	30,794
Low Metascore	0.018	0.133	0	1	30,794
High Metascore	0.221	0.415	0	1	30,794

Table 1.5: Demand Model Marginal Utilities

Variable	Marginal Utility	95% Lower	95% Upper
Independent	-0.786***	-1.052	-0.520
	(0.136)		
Free	-0.865***	-1.112	-0.618
	(0.126)		
Price	-0.001***	-0.002	-0.001
	(9.86×10^{-5})		
Price Std. Dev.	1.31×10^{-16}		

*p<0.1, **p<0.05, ***p<0.01

Standard errors are in parentheses.

Table 1.6: Demand Model Marginal Utilities for Reviews

Variables	Marginal Utilities For Independent	Marginal Utilities For Major
Score Rank	0.011***	0.0249***
	(0.003)	(0.004)
Score Rank^2	0.033×10^{-4}	-1.802×10^{-4}***
	2.811×10^{-5}	(3.33×10^{-5})
No Metascore	-0.418***	-0.565***
	(0.064)	(0.073)
Low Metascore	0.333	-0.228
	(0.239)	(0.204)
High Metascore	0.334***	0.838***
	(0.081)	(0.087)

*p<0.1, **p<0.05, ***p<0.01

Standard errors are in parentheses.

APPENDIX B: CHAPTER I FIGURES

Note: This is the top of the game's page.

Figure 1.1: User Score

Note: This is a short ways distance from the top of a game's page.

Figure 1.2: Metacritic Score

CHAPTER II

NEWS CONSUMPTION IN A SOCIAL MEDIA ENVIRONMENT: EVIDENCE FROM TWITTER

1 Introduction

Individuals are consuming an increasing amount of news and information through social media. According to a 2017 Pew Research Center Survey, 67% of US adults consume news on social media and 20% do so often (Shearer and Gottfried 2017). Social media presents a platform in which individuals may be exposed to a variety of news sources. These sources differ in terms of their political ideology and their accuracy. Because of opportunity costs and preferences, individuals likely do not consume all news they come across. This article seeks to shed light on what drives individuals' consumption decisions concerning news on social media. In particular, the study will focus on preferences for ideology and for accuracy in news. In other words, this work will begin to shed light on the demand for news in a social media environment as it concerns ideology and accuracy.

I examine the importance of ideological similarity between news sources and individuals and the importance of the reliability of a news source in US individuals' consumption decisions in a social media environment by estimating a simply utility function. I examine this preferences separately for liberal individuals and conservative individuals. In particular, I ask two questions: 1) Do consumers have a preference for news with a similar or opposing ideological position to their own? I find, conditional on individuals' following decisions, liberal individuals have an increasing preference for news as it becomes more conservative (i.e. less similar) and no preference for news as it becomes more liberal. I find conservative individuals do not have preferences for more or less similar ideological news. These findings may be the result of the timing of data collection. During a time when liberals can be thought of as the minority party, they may consume conservative news, particularly on social media, to mock what they view as an incorrect viewpoint. 2) Do consumers have a preference for accurate or inaccurate reporting? The results indicate liberal individuals prefer less accurate news, which is likely driven by sensationalism. Conservative individuals have no preference for more or less reliable news. Understanding the consumption behavior of individuals on social media is of interest to economists. Consumer preferences, as a part of utility, are an important component of demand. In particular, economists wishing to model the market for news in a social media environment need a strong understanding of consumer demand in that market. In the past economists have built models of the market for news in other environments and the current work can shed light on if past assumptions about consumer utility apply to the news market on social media (see Mullainathan and Shleifer 2005; Gentzkow and Shapiro 2006; Allcott and Gentzkow 2017). Understanding the consumption habits for individuals on social media is also of interest to news organizations, advertisers, and social media companies as these preferences will affect their behavior. This is especially true as social media becomes a more important market for news.[1]

This work is the first empirical study, to my knowledge, to examine the effect of both ideological preference and reliability on news consumption decisions in a social media environment using data on individuals' ideology, news' ideology, and news' reliability. There

[1] From 2016 to 2017 there was a 15 percentage point increase in the number of Twitter users consuming news from the platform (Shearer and Gottfried 2017).

are several studies closely related to my own. Boxell, Gentzkow, and Shapiro (2017) find political polarization in recent years is largest in demographic groups least likely to use social media. Their findings argue against the idea the internet and social media are primary drivers of polarization. Gentzkow and Shapiro (2011) find low levels of ideological segregation, the situation where conservatives consume conservative news and liberals consume liberal news, in individuals' online consumption habits. They develop a structural model estimating the effects of ideology and quality on consumer utility to test their assumptions (Gentzkow and Shapiro 2011). While similar, there are several differences between their work and the current study. First, while quality is related to my concept of reliability it is not identical. I define reliability to be the likelihood of factual reporting from a news source. While factual reporting is a component of what Gentkow and Shapiro refer to as quality, quality includes other features such as entertainment, timeliness, and writing standards. Another difference in their work is quality is a parameter to be estimated. In my work I use data on reliability and estimate the effect of reliability on consumption decisions. This implies I do not assume a priori reliability is a feature which gives consumers positive utility. In other words I do not assume consumers view factual reporting as resulting in higher quality.

Another related study, An et al. (2014) looks at the sharing of news articles on Facebook and Twitter. The study finds that partisan sharing, the situation where individuals share news similar to their own beliefs, exists on social media. Finally, Allcott and Gentzkow (2017) look at fake news consumption on social media and the likelihood of users to believe what they consume. They estimate that the average US adult read and remembered about one fake article.

The remainder of the paper is organized as follows. Section 2 gives a brief description of Twitter. Section 3 of the paper presents a model of utility for news consumption on social media.

Section 4 describes the data and sources. Section 5 introduces the empirical strategy. Section 6 presents the findings and possible explanations for the results. Section 7 concludes.

2 Twitter

Twitter is a social media platform which allows users to create accounts and post and view short messages called tweets. The first thing a user encounters upon logging into their account is what is called their home timeline. An account's home timeline is populated by tweets from friends. Friends are accounts an individual has decided to follow. Tweets are sorted into several categories. A standard tweet is an original message posted by an account. A retweet is a reposting of another account's tweet verbatim. A quote tweet is the same as a retweet except a small message is added to the original tweet by the reposting account. Another type of tweet which may enter an account's home timeline is a reply. A reply occurs when one account directly responds to another account's tweet. Replies will enter a particular account's home timeline only if the host account of the home timeline is friends with both the account responsible for posting the original tweet and the account posting the reply.

To attempt to keep explanations clear, I refer to accounts an individual chooses to follow as friends and accounts which choose to follow an individual as followers throughout this study.[2]

Content Twitter believes is relevant to an account may be posted to the top of their home timeline at the time an individual logs into their account. This is determined by an algorithm and includes popular tweets from friends of accounts and, sometimes, tweets from non-friends. The

[2] For additional information on Twitter see the Twitter Help Center (2018) and the Twitter Blog (2018).

rest of the activity populating a home timeline consists of tweets from friends posted in reverse chronological order (Oremus 2017).[3] The relative simplicity of most of Twitter's home timeline creates a social media environment easier to study than many other social media platforms, such as Facebook, which uses more complex algorithms to create what users view.[4]

Twitter also provides account's access to their user timeline. This is a timeline which is populated by the tweets, retweets, quote tweets, and replies produced by the account. The user timeline is a record of an account's activity excluding the favoriting of tweets, which is an activity similar to liking a post on Facebook.

Another feature of Twitter in need of some discussion is how Twitter classifies accounts. Twitter classifies accounts as either verified or non-verified. Verified accounts are defined as "accounts of public interest" including "accounts maintained by users in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas" (Twitter Help Center 2018). Essentially, verified accounts are those maintained by high profile individuals or businesses. This tag is meant to allow individuals to determine if an account is run by the person who claims to manage said account. My study focuses on non-verified accounts as verified accounts likely have different behavior than accounts operated by less high profile individuals.

Twitter is a growing source of news for many. 74% of individuals active on Twitter and 11% of the US adult population get some news from the site (Shearer and Gottfried 2017). This suggests Twitter is an important part of the news market and an important market to study.

[3] This is the most up to date information I could find about the Twitter home timeline. A detailed discussion on the topic of Twitter home timelines can be found in Oremus (2017).

[4] Facebook uses a complex algorithm to choose all the content users view on their home feed (Oremus 2017).

3 A Utility Model of News Consumption on Social Media

Consider a representative rational agent i, with an active account[5] on social media, who has ideology b. At time t a unique reporting of news from source j with ideology a, reliability R, and price P enters the agent's home timeline. Agent i chooses whether or not to consume the news based on the following utility function:

$$u_{ijt} = \alpha + \mu_i + \beta(a_{jt} - b_i)Right_{ijt} + \theta(a_{jt} - b_i)Left_{ijt} + \gamma R_{jt} - \lambda P_{jt} + \epsilon_{it}, \quad (1)$$

where u_{ijt} is the utility derived from consuming news. I assume the utility from not consuming news is normalized to zero. The agent will choose to consume news if they receive any positive utility from doing so.

α is a term representing utility common to all individuals. μ_i is an individual specific term which can be viewed as an individual's utility or disutility from consuming news or the individual preference for news. a_{jt} is the ideology of source j producing the news which enters agent i's home timeline at time t. b_i represents agent i's ideology and is constant. I define ideology as a unidimensional measure of political preference for both individuals and news outlets standardized to have a mean of zero.[6] The difference $(a_{jt} - b_i)$ captures how similar j's ideology is to agent i's ideology. I separate this distance into two terms to allow agent i to have varying preferences for news to the right and news to the left. The term $(a_{jt} - b_i)Right_{ijt}$ takes on the value of $(a_{jt} - b_i)$ if j's ideology is to the right of i's ideology and zero otherwise while the term of $(a_{jt} - b_i)Left_{ijt}$ takes on the value of $(a_{jt} - b_i)$ if j's ideology is to the left of i's

[5] I assume active accounts are similar to the accounts examined in Barberá et al. (2015).
[6] My definition of ideology is the same as that in Barberá et al. (2015).

ideology and zero otherwise. β and θ capture the preferences for ideological similarity of news to the right and to the left of agent i, respectively. Some care must be taken with interpreting the coefficients. If agent i's utility increases (decreases) when source j is more ideologically similar to i (i.e. $(a_{jt} - b_i)$ is smaller) and news source j's ideology is to the right of i's ideology, β should be negative (positive). However, if same situation occurs when news source j's ideology is to the left of agent i's ideology, θ should be positive (negative).[7]

R_{jt} is a measure of the reliability of news from source j which enters agent i's home timeline at time t. I define reliability as the likelihood a news source produces an accurate report of events. R_{jt} takes on a value of one, two, or three with one representing the lowest level of reliability (higher likelihood of reporting a false report) and three representing the highest level of reliability (lower likelihood of reporting a false report). Reliability is distinguishable from truth because, even if a source consistently seeks to report true information, mistakes can be made. γ represents the preference for reliability. γ will be positive if agent i receives positive utility from reliability and negative if they receive negative utility.

P_{jt} represents the price of consuming news from source j at time t and λ captures the effect of price on utility. λ is subtracted from utility as it is assumed higher prices reduce utility. I assume P_{jt} is zero.[8] It is beyond the scope of this study to present a detailed discussion as to why price is zero for news on social media, however, a simplified explanation is provided. The market for news on social media is an example of a two-sided market.[9] A two-sided market is a

[7] $(a_{jt} - b_i)Right_{ijt}$ will always be positive and $(a_{jt} - b_i)Left_{ijt}$ will always be negative or zero.

[8] There are a few sources, such as the Economist, which only allow access to a limited number of articles per week without a paid subscription. However, these cases are rare and it is not expected this has a great impact on social media news consumption.

[9] It actually could be considered an intersection of several two-sided markets.

market in which two groups interact through an intermediary and the actions of each group have an effect on the outcomes of the other group. For example, in the social media market consumers are one group and advertisers are the other. These groups interact through the social media platform (e.g. Twitter). Economic theory suggests pricing to both groups in a two-sided market depends, in part, on the price elasticity of demand for both sides of the market. If one side of the market has a high price elasticity of demand relative to the other side of the market, the side with a higher price elasticity of demand may have a zero or even negative price (Rysman 2009). On social media, advertisers (with more inelastic demand) essentially subsidize the participation of social media users (with more elastic demand). The same is often the case in the news market where advertisers, again, subsidize individuals' consumption of the news. This is a simplified explanation of the news market on social media, however, it helps illuminate the reason for the zero price faced by agent i in the social media news market. This theory is why I drop price from (1) and allows the utility function to be rewritten as:

$$u_{ijt} = \alpha + \mu_i + \beta(a_{jt} - b_i)Right_{ijt} + \theta(a_{jt} - b_i)Left_{ijt} + \gamma R_{jt} + \epsilon_{it}. \quad (2)$$

Finally, ϵ_{it} it is idiosyncratic utility. Note t does not index a specific time but rather the order news enters the home timeline.

It should be noted the news which enters the home timeline is not random. It depends on factors such as who the agent chooses to follow. I assume the decision to follow someone and the decision to consume news are separate. As long as the agent does not consume every piece of news sent by their friends this assumption seems reasonable. Additionally, I assume each reporting of news entering an account's home timeline is unique.[10]

[10] This assumption is made for practical reasons. The assumption implies each piece of news can only be considered for consumption once. Without this assumption, the same news could enter

It is assumed agent i has full knowledge of their preferences and the values of a_{jt}, b_i, R_{jt}, α, and μ_i at time t. Agent i uses this information at time t to decide whether to consume news from source j or not. I make no assumption about ideology being a negative quality. Instead ideology is simply a characteristic of the news. News can be presented which is both ideological and reliable.

Estimating the utility function will provide insight into consumer preferences, an important component of demand, for the news market on social media.

4 Data

4.1 Twitter Sample

The first step in collecting the data for estimation is constructing a sample of active US Twitter accounts. I do this using Twitter's API service (Twitter Developer 2017).[11]

I drop 'deleted' tweets and apply a filter suggested by Barberá et al. (2015) intended to be a simple location and activity filter.[12] I remove all accounts listed as verified to remove accounts

an account's home timeline multiple times and, even if the news is consumed, count as non-consumption multiple times.

[11] I collect this sample over the period from June 20 to June 22, 2017. To primarily collect US accounts I open a connection to Twitter's Streaming API from 11 a.m. to 3 p.m. Central Time each day. This gives a random sample of roughly 1% of all public tweets. The specified time period coincides with noon to 1 p.m. for each of the four time zones in the continental US. This is the peak time for Twitter activity in each time zone (Lee 2016). A potential problem is the time period when the sample is collected differs across time zones possibly biasing the sample. I expect that bias is minimized since the majority of tweets collected should come from accounts in the most active time zone at any given hour.

[12] The filter drops all accounts who do not identify English as their language, who have tweeted less than 100 times, who have less than 100 friends and 25 followers. This filter is used because while some accounts have location data it is not consistent or reliable.

related to a news source and, additionally, to remove accounts run by high profile individuals who may behave differently on Twitter and are not the population of interest for this study.

From the tweets I am able to collect a sample of 530,416 unique accounts. Twitter's API service (Twitter Developer 2017) rate limits restrict my ability to work with such a large sample. I randomly draw 1,500 accounts to create a sample for which data can be obtained in a reasonable time period.

4.2 User Timelines and Home Timelines

To estimate the parameters of my model, it is necessary to observe both potential and actual consumption for each account. An account's home timeline cannot be directly observed. However, I am able to construct a partial approximation of an account's home timeline using information about an account's friends and an account's user timeline. I gather the list of friends for each account in my sample.[13] Accounts which have been suspended, become private, or have been deleted during the time between gathering data on random tweets and data on accounts' friends are dropped leaving me with 1,412 users. Twitter API rate limiting makes it unreasonable to collect tweets from every friend of accounts in my sample.[14] To address this issue, I draw 80 friends from a weighted random sample, where friends are weighted by the total number of tweets sent since activated. This has the effect of drawing 80 friends at random with a sampling

[13] I collect this data from Twitter's API (Twitter Developer 2017) between September 22 and 23, 2017.
[14] Accounts in the sample follow up to 70,955 accounts.

method which approximates the likelihood a friend sent tweets which show up in an account's home timeline.[15]

I collect the last 200 tweets from accounts in my sample and the random sample of their friends.[16]

Some accounts and friends may have less than 200 tweets. This is caused by deleted tweets and the fact some accounts have not sent 200 tweets since activated. This does not give me a perfect recreation of an account's home timeline. Instead, I am able to create a partial home timeline for each account. This should not be a major concern because the estimation procedure does not require I observe an account's entire home timeline for the estimates to be valid.[17]

4.3 News Sources

To construct a set of news sources, I take the list of news sites used to estimate the structural model in Gentzkow and Shapiro (2011) and remove any sources for which there is no verified Twitter account.[18] For the remaining sources I augment the collection of news sources by adding accounts to some of the outlets which have multiple Twitter accounts. For example, I include the

---

[15] A few accounts in my sample have less than 80 friends. While the filter applied when collecting the sample implies this should not be possible there are some rational explanations for why I obtain this result. Two prime examples are accounts may choose to remove certain friends and some friends may be deleted or suspended thus decreasing the total number of an account's friends.

[16] I collect 200 tweets because it is the maximum number of tweets which may be collected in a request. These 200 tweets include all tweets on their user timeline which have not been deleted.

[17] The tweets were collected between October 6 and October 15, 2017. The time period between collecting data on accounts' friends and collecting accounts' and friends' user timelines can be thought of as a spam filter. Twitter accounts that are active over long periods are unlikely to be spam bots (Barberá et al. 2015).

[18] See their Model Appendix for a complete list.

FoxNews, foxheadlines, and foxnewspolitics accounts. Finally, news source without an estimate for ideology and without data on reliability, discussed in sections 4.6 and 4.5 respectively, are dropped leaving me with 44 Twitter news accounts. I augment these accounts once again by including 11 news accounts without an ideology estimate but which come from the same parent organization as an account for which I do have an ideology estimate.[19] This leaves me with 55 news accounts. My final list of news sources, along with their reliability, ideology, and number of followers is given in Table 2.1.

4.4 Consumption

It is difficult to directly observe what news individuals on Twitter consume. However, I am able to use the given data to create a proxy for consumption and potential consumption. The list below presents the actions which I consider to represent consumption.

1. A retweet of a news source.
2. A quote tweet of a news source.
3. A retweet of a quote tweet of a news source.
4. A reply to a news source.
5. A retweet of a reply to a news source.
6. A quote tweet of a reply to a news source.
7. A retweet of a quote tweet of a reply to a news source.
8. A quote tweet of a quote tweet of a news source.

[19] For example, I give foxheadlines the same ideology score as FoxNews.

9. A retweet of a quote tweet of a quote tweet of a news source.

Retweets, and other similar actions on Twitter, are signals of engagement with a tweet. This proxy for consumption is used because it captures an account's engagement with information from a news source. I observe these actions from an account's user timeline.

Combining my assumption about the actions which constitute consumption with the constructed partial home timelines, I am able to identify the consumption and non-consumption of news by accounts in my sample.[20]

Potential consumption, which is all news considered for consumption by an account (i.e. the sum of consumption and non-consumption), may come directly from a news source or indirectly through another friend.[21] This allows accounts in my sample to potentially be exposed to a more diverse set of news.[22] I am left with a sample of 331 accounts who potentially consume information from a news source in Table 2.1.

Note my definition of consumption does not consider consumption to be clicking on an article. This is because information can often be consumed from the content of a tweet. Figure 2.1 gives an example. Objective breaking news headlines are another good example of tweets which can be consumed without clicking on a link. Capturing this type of consumption is one advantage of my proxy over using link clicks. Another advantage of my definition of

---

[20] Favorites are not considered consumption because they require the least amount of input out of all actions Twitter permits accounts to make. Favorites are, therefore, much less likely to signal any significant engagement with the content of the tweet. Additionally, I only allow a unique Tweet ID to be considered for consumption once by each account as assumed by the model in Sec 3. I also drop any reply tweets which were not in response to an account's friend and therefore would not enter that particular account's home timeline.

[21] For example, a friend who retweets Fox News.

[22] This is justified based on the work of An et al. (2011) who find indirect media exposure, exposure to news which does not come directly from the source, increases the ideological diversity of news individuals are exposed to by a significant margin.

consumption is data is available.[23] The definition of consumption used for this article is certainly reasonable if it is assumed consumption requires a certain level of engagement with the content.[24]

4.5 Reliability

I need data on the reliability of news sources. What is true is often difficult to determine. As mentioned above, I define reliability as the likelihood of factual reporting from a news source. My definition of factual reporting is the production of an accurate report of events. However, it is perhaps better to think about what is not factual reporting. I consider non-factual reporting to be the case of reporting events which are verifiably false, not cases where groups with different points of view may disagree about the accuracy of the information.

The data on news' reliability comes from MediaBias/FactCheck (2017). This source has several advantages. First, it gives each news source a Factual Reporting score. These scores are partially assigned based on research into how many fact checks a source has failed. In particular, the site references PolitiFact and Snopes frequently in their assessments of Factual Reporting. Both PolitiFact and Snopes have been used in academic research on fake news (see Allcott and Gentzkow 2017).[25]

---

[23] I do not have information on who clicks what link.

[24] If engagement with a tweet is not considered a good proxy for consumption, the findings of this study would instead reflect preferences in promoting behavior. In other words, the results would give insight into what news individuals prefer to spread to their followers.

[25] MediaBias/FactCheck (2017) has also been be suggested as a reference to help assess either the reliability or bias of a news source by several academic institutions' libraries. Examples include the University of Minnesota Libraries (2017), the University of Pittsburgh: University Library System (2017), and the University Libraries: University of Washington (2017).

The Factual Reporting score places the news source in one of 5 bins (Very Low, Low, Mixed, High, and Very High). The higher the factual reporting score the more likely the source is to not have failed many fact checks. All the news sources in my list have a score of mixed or higher and are coded on a scale of one to three with one being mixed and three being very high. The exclusion of Low and Very Low news sources is the result of having no ideology estimates from Barberá et al. (2015). The reliability score is assigned based on a news account's parent organization.[26]

4.6 Ideology

Agents' and news sources' ideology are not observed directly.[27] My data on ideology is derived from Barberá's work developing a method to estimate the ideology of agents and political ‘elites' on Twitter (See Barberá 2015; Barberá et al. 2015). Elites are accounts such as political actors, news organizations, and think tanks for which ideology is a key component of an agent's decision about whether to follow the elite account. Barberá (2015) and Barberá et al. (2015) develop two versions of an ideal points model to estimate ideology, a unidimensional variable, for both agents and elites in a multi-step process. The key identifying assumption for both

[26] To further validate MediaBias/FactCheck (2017) I test the correlation between the overall percentage “Pants on Fire” scores for five major channels on PunditFact (2017) (sister site to PolitiFact) and my 1 to 3 score based on MediaBias/FactCheck (2017). The 5 sources for which PunditFact (2017) scores a significant amount of statements are ABC, CBS, NBC/MSNBC, CNN, and Fox. For the NBC/MSNBC score from MediaBias/FactCheck (2017) I use the average of the two scores. The Pearson's correlation is -0.878 which suggests that a higher score based on MediaBias/FactCheck (2017) is associated with a lower percentage of major fact check fails from PunditFact (2017). However, it should be kept in mind this is a correlation of only 5 observations so it is at best suggestive evidence.

[27] Here the agent is representative of the users of accounts which are the focus of this study.

models are agents prefer to follow accounts, particularly political and news accounts, which are similar to their own ideology and thus following decisions are informative about agents' and elites' ideology. I choose to use estimates from the model in Barberá et al. (2015). The advantage of this version of the model is an additional stage which expands the number of elites and allows for the estimation of ideology for a larger number of agents (Barberá et al. 2015). The authors model the probability an agent decides to follow an elite account as function of the euclidean distance between the elite's and agent's ideology, the popularity of the elite account, and the political interest of the agent.[28] The greater the distance between the ideology of an agent and an elite, all else equal, the less likely the agent is to follow the elite. Given the decision to follow an elite account is binary, the authors assume the error term is logistically distributed. Using correspondence analysis, which is similar to a log-linear latent space model, Barberá et al (2015) are able to estimate the parameters of their model of following behavior. These parameters allow them to obtain estimates of the ideology of agents and elites. This is done in three stages. The first stage estimates the model parameters for elite accounts and for agents which follow at least 10 elites. Once estimates for the ideology of these agents and elites are obtained, the second stage identifies the most popular accounts followed primarily by liberal agents and the most popular accounts followed primarily by conservative agents in the first stage and labels them elites.[29] This is based on the assumption following one of these popular accounts, even though they are not necessarily political, provides information about an agent's ideology. The third stage

[28] Using the notation of Barberá et al. (2015) the probability agent i follows elite j is formally expressed as $Pr(Y_{ij} = 1 |\ \alpha_i,\ \beta_j, d_{ij}) = Logit(\alpha_i + \beta_j - d_{ij})$ where α_i captures agent i's political interest, β_j captures the popularity of elite j and d_{ij} is the Euclidean distance between agent i's ideological position and elite j's ideological position. The function is designed such that the further the distance between the ideology of i and j the less likely i is to follow j. This notation should not be confused with the notation used throughout the paper.
[29] Liberals and conservatives are defined based on their estimated ideology.

estimates the model using the expanded number of elites allowing a larger number of agents to be incorporated in the analysis.

The ideology estimates are standardized to have a standard normal distribution (Barberá et al. 2015). More negative (positive) values are associated with more liberal (conservative) ideology. Barberá (2015) validates the original model results for elites by comparing them to Poole and Rosenthal (2007) DW-NOMINATE measures of ideology, and the results for agents by successfully matching a sample of agents to party registration records and a sample of agents to publically stated political preference.[30] The Barberá (2015) and Barberá et al. (2015) estimates for ideology are highly correlated (Barberá et al. 2015).

Instead of estimating the full model myself, I use the tweetscores package in R (Barberá 2016) which eliminates the need to estimate the first two stages of the Barberá et al. (2015) version of the model. This saves an immense amount of time as it eliminates the need to collect data on all followers of elites. All I need to estimate ideology for an agent with an account in my sample is the list of their friends. The estimates are based on data gathered in 2014. I do not expect this invalidates the estimates as it is unlikely large swaths of politicians and news organizations have dramatically shifted their ideological positions over the intervening period.[31] In addition to the ideology for agents with accounts in my sample, I need ideology estimates for my list of news sources. I use the Barberá et al. (2015) ideology estimates for elites and match them with the list of verified news accounts.[32]

[30] The results are also validated in several European nations.
[31] Only accounts which follow at least one elite remain in sample.
[32] This is extracted from the tweetscores package (Barberá 2016).

4.7 Descriptive Statistics

Table 2.2 breaks down the sample into liberal (ideology ≤ 0) and conservative (ideology > 0) accounts. Liberal accounts outnumber conservative accounts roughly 60 to 40. Recall the sample is of active US Twitter accounts. If liberal accounts are more likely to be active on Twitter this would explain why there is not a more even division between liberals and conservatives.[33]
It should be kept in mind Twitter is likely not representative of the US population at large and therefore liberals and conservatives on Twitter may not be a perfect match for liberals and conservatives elsewhere.

Table 2.3 shows consumption for accounts in the sample. News consumed is coded as one while news not consumed is coded as zero. It is immediately apparent accounts consume a low amount of news relative to the amount of news they see. This is not an entirely surprising result as accounts are confronted with an enormous amount of information online and consuming a large percentage of it would have high opportunity costs. Table 2.4 shows consumption for liberal and conservative accounts and tells pretty much the same story as Table 2.3. Both liberals and conservatives consume a small percentage of total potential consumption.

Figure 2.2 presents a histogram of the frequency of the reliability measurement for news accounts and Table 2.5 gives summary statistics. More news sources fall into the high category than any other category while relatively few sources fit in the very high category. Figure 2.2 may suggest agents are primarily exposed to news from the mixed and high categories and rarely exposed to news from the very high category.

[33] Barberá also finds liberals are a majority in his sample (Barberá 2015).

Table 2.6 shows the number of accounts, consumption, potential consumption, and consumption rates for each reliability category.[34] The information presented in Table 2.6 supports the notion accounts are exposed to a high volume of news from sources in the high category. Perhaps the most striking information in Table 2.6 is how much potential consumption and consumption there is in the very high category despite there only being seven news sources in that category. This alleviates some of the concern accounts are rarely exposed to news from the most reliable news sources. Also of note is while accounts are exposed to more information from the high category, they consume a smaller percent of potential consumption from these sources. Additionally, accounts consume a higher percent of potential consumption from the mixed category than from any other group.

Table 2.7 and 2.8 shows consumption, potential consumption, and consumption rates by reliability for liberal accounts and conservative accounts respectively. Liberals are exposed to a higher percentage of high reliability tweets, 65% of total potential consumption, than their conservative counterparts, 43% of total potential consumption. On the other hand, conservatives are exposed to a much higher percentage of mixed tweets, 38% of total potential consumption, than liberals, 18% of total potential consumption. Both liberal accounts and conservative accounts are exposed to roughly the same fraction of very high tweets, 17% and 19% respectively. This highlights how the exposure to reliability varies between conservatives and liberals. This exposure is a function of the friends an account chooses which is not modeled in the current paper. It should also be noted both liberals and conservatives consume a higher percent of potential consumption from news sources in the mixed category than from the other news sources.

[34] Potential consumption is the sum of consumption and non-consumption.

Figure 2.3 presents the density distribution of news outlets' ideology. There is a very large peak left of zero and a smaller peak right of zero. This indicates the news sources used for this study are primarily liberal (ideology ≤ 0) with a small group of conservative (ideology > 0) sources. The information from Table 2.9 further supports this notion. Over 85% of all news sources in the sample are liberal. Table 2.5 shows the average news' ideology is somewhat left of zero.

Table 2.10 shows consumption, potential consumption, and consumption rates for the full sample while this information is broken down for liberal accounts in Table 2.11 and conservative accounts in Table 2.12. For the full sample, conservative news is makes up 9.85% of accounts' exposure to news while liberal news is responsible for the other 90.15%. The difference in exposure is even more dramatic for liberal accounts. Conservative news only represents only 2.12% of total exposure to the news for liberal accounts. On the other hand, 24.73% of conservative accounts' exposure to the news comes from conservative news sources. The fact conservative news sources are responsible for a higher percentage of conservative accounts' exposure to the news than for liberal accounts and vice versa is not surprising and highlights the importance of accounts' decisions about who to follow.

It is interesting to note even though accounts are exposed to a greater amount of liberal news, they consume a higher percentage of the conservative news they encounter. Liberal accounts in particular have a much higher consumption rate for the conservative news which comes through their home timeline relative to their consumption rate for liberal news. The difference in the consumption rates of liberal and conservative news for conservative accounts is smaller.

Figure 2.4 gives a box plot graph of news' ideology by news' reliability. The first thing which stands out is there is more variance in the ideology of news sources with mixed reliability than news sources in the other two categories. Similarly, news sources with high reliability have more variance in ideology than news sources in the very high category. The information in Figure 2.4 suggests for many accounts there may be a tradeoff between ideology and reliability. That is, one may have to consume news which does not match their ideology in order to consume reliable news.

It should be kept in mind, when examining my data related to news sources, the sample of news accounts is not representative of every news source on Twitter. Therefore, caution must be taken when generalizing the findings to news sources as a whole.

5 Empirical Methodology

While utility, u_{ijt} is unobserved, I can estimate equation (2) using data on consumption. Because I observe whether news is consumed or not, I can define consumption, c_{ijt}, such that:

$$c_{ijt} = \begin{cases} 1, if\ u_{ijt} > 0 \\ 0, if\ u_{ijt} = 0 \end{cases}, \quad (3)$$

allowing me to estimate the utility model as a binary choice model. Because I have a panel data set, using proper assumptions, I can estimate the marginal effects of ideological distance of news with ideology left of an account, ideological distance of news with ideology right of an account, and reliability of a news source on the probability of consumption by using a random effects probit model.

The first assumption which must be made to run the model is that $\epsilon \sim N(0,1)$ and the individual specific component of utility $\mu \sim N(0, \sigma_\mu^2)$. An additional assumption which must be made to identify the random effects probit model is that μ_i, the individual preference for information, is uncorrelated with $(a_{jt} - b_i)Right_{ijt}$, $(a_{jt} - b_i)Left_{ijt}$, and R_{jt}. It is likely a preference for information is uncorrelated with either ideological distance or reliability. The higher one's preference for information the more likely an individual is to consume information regardless of the ideological content or the reliability of the news source because they simply like to consume information. Therefore, it is fairly safe to assume μ_{ijt} is uncorrelated with $(a_{jt} - b_i)Right_{ijt}$, $(a_{jt} - b_i)Left_{ijt}$, and R_{jt} .

I choose to model the utility with a random effects probit model because of the ability to allow for unobserved individual heterogeneity. In a market for news it is probable there will be a great deal of unobserved differences between individuals' preferences for news. Some individuals likely have a strong taste for information and consume a great deal of news while other individuals may be very casual consumers. In the random effects probit model heterogeneity enters the model through the term μ_i or what I have called the preference for information. This term varies across individuals by allowing each individual to deviate from the mean value.[35] The term μ_i allows individuals to have different preferences for news on social media.

I run the utility model specified in (2) separately on liberal accounts and conservative accounts. This allows me to see if consumer preferences vary for groups who potentially have different tastes. Conventional wisdom suggests differences in tastes between liberals and conservatives may be particularly pronounced when examining presences for ideology.

---

[35] This can be thought of as deviation from the constant value α.

6 Results and Discussion

The estimated marginal effects of the model specified in equation (2) for liberal accounts are presented in Table 2.13. The model indicates the proportion of total variance contributed by panel-level variance is not zero, therefore the panel probit is preferred to a pooled probit model. In terms of the utility model, this implies liberals have different preferences for news.

The results for liberals indicate the marginal effect on ideological distance to the right is positive and statistically different from zero. This suggests liberals are more likely to consume news as it becomes more ideologically conservative.[36]

The marginal effect on ideological distance to the left is not statistically different from zero suggesting liberals have no preference for news whose ideology is more liberal than their own. The marginal effect for news' reliability is negative and statistically significant at the 5% level. That is, increasing the likelihood of factual reporting from a news source decreases liberals probability of consuming news from that source.

Table 2.14 presents the estimated marginal effects of the model specified in (2) for conservative accounts. Like the model for liberal accounts, the proportion of total variance contributed by panel-level variance is not zero and the random effects probit is preferred to the pooled probit.

The results for conservatives are quite different from the results for liberals. The marginal effects on ideological distance to the right, ideological distance to the left, as well as news' reliability are all not statistically different from zero for conservatives. This suggests

[36] Conservative here is relative. A news source to the right of a liberal account's ideology is not necessarily conservative (i.e. ideology>0).

conservatives have no preference for ideology, whether the news source is to the right or to the left. Additionally, conservatives do not have a preference for more reliable or less reliable news.

While at first the finding liberals prefer less reliable news may seem surprising, there is a plausible explanation for the findings. It is likely less reliable news implies more sensational news. To test the notion less reliable news is correlated with more sensational news I collect additional data from MediaBias/FactCheck (2017). News sources are divided into categories such as “Least Biased”, “Left-Center/Right-Center Bias”, and “Left/Right Bias.” A significant factor in sorting news sources into these categories is their use of loaded words which the website describes as “wording that attempts to influence an audience by using appeal to emotion or stereotypes” (MediaBias/FactCheck 2017). I use this information to code news sources on a scale of one to three with one being news in the Least Biased category and three being news in the Left/Right Bias category giving me a variable measuring the degree of sensationalism of a news source. A larger value implies a more sensational news source while a smaller value implies a less sensational news source. I find the Pearson's correlation between reliability and sensationalism to be -0.678. This supports the idea less reliable news is more sensational news. Given less reliable news is more sensational news, the results imply liberals have a preference for sensational news over more objective and accurate news. It is plausible liberal accounts on social media, where an enormous amount of information is confronted, would consume the news which is the most shocking and dramatic. Conservatives, on the other hand, have no preference for more reliable or less reliable news. For conservative accounts, it seems, reliability does not factor into the consumption decision. It should be kept in mind this study only uses news accounts above a certain level of reliability. The preference for reliability may change if low reliability or fake news accounts were included in the estimation.

The finding liberals have a preference for news which is more conservative than their own ideological position and not for news which is more liberal than their ideological position is unexpected but is not unreasonable. One potential explanation is liberals enjoy mocking viewpoints they consider incorrect, especially at a time when individuals with liberal ideological views do not control the government. These results may be driven by the timing of the data collection. Given the data was collected at a time when Republicans control the Executive and Legislative branches of government, liberals may be keen on consuming and deriding news from more conservative sources.[37] This explanation seems likely when considering liberals have a preference for less reliable, and sensational, news. If liberals are consuming more conservative news in a mocking manner it seems likely the news they are consuming would be less authentic and perhaps an easier target for ridicule. Conservatives have no preference for more liberal or more conservative news. Perhaps conservative, in a time when they may be considered the majority party, consume a fairly ideologically balanced diet of news. Again, this explanation is tied to the timing of the data collection. It should be kept in mind the current study only reflects the preferences of individuals on Twitter. It is entirely possible agents consume news differently outside of social media (e.g. liberals consume liberal news on TV but conservative news on social media). It would be interesting to see if the findings were reversed during a period when Democrats controlled the White House and congress.

It is important to keep in mind the findings do not tell us anything about if individuals believe what they consume. Just because liberal individuals may have a preference for less reliable, and more sensational, news does not imply they believe or do not believe the contents of

[37] Not all tweets necessarily occur during the period between October 6 and October 15, 2017, however, since the sample is constructed from the last 200 tweets starting from this time period and moving backwards in time, it is highly likely the tweets of active accounts are recent.

what they consume. Additionally, liberal individuals do not necessarily become more conservative when they consume conservative news. Finally, it is important to keep in mind the results are conditional on who individuals choose as friends on Twitter. That is, these preferences do not reflect random articles entering a home timeline. This should not be a large concern, however, as news does not enter an account's home timeline in reality.

7 Conclusion

This study has set out to examine consumer preferences for news in a social media environment. In particular, I set out to answer two questions. First, do consumers have a preference for news with a similar or opposing ideological position to their own? My findings suggest liberals have a preference for more conservative news while they have no preference for more liberal news. Conservatives, on the other hand, have no preference for more conservative or more liberal news. This may be a function of gathering the data in a year when the liberal party is the minority party in Washington implying liberals gain some utility from mocking conservative views when individuals with such views are in power while conservatives gain no such utility at the same time. Second, do consumers have a preference for accurate or inaccurate reporting? Liberals prefer less reliable news. Conservatives have no preference for more or less reliable news. These results suggest liberals may enjoy consuming more sensational news while conservatives do not have similar preferences.

It should be kept in mind all of the above findings are conditional on who individuals choose to follow and the findings say nothing about individuals' beliefs about what they consume. However, the results still provide important insights for economists wishing to have a better grasp of consumer demand in the news market on social media. Understanding news consumption habits is also of interest to a variety of groups including news producers, advertisers, and social media platforms who may adjust their behavior in the market for news on social media as it continues to grow.

REFERENCES

1. Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." Journal of Economic Perspectives 31(2):211-236.

2. An, Jisun., Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2011. "Media Landscape in Twitter: A World of New Conventions and Political Diversity." Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

3. An, Jisun. Daniele Quercia, and Jon Crowcroft. 2014. "Partisan Sharing: Facebook Evidence and Societal Consequences." COSN '14 Proceedings of the second ACM conference on Online social networks.

4. Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation." Political Analysis 23(1):76-91.

5. Barberá, Pablo. 2016. "twitter_ideology." GitHub. https://github.com/pablobarbera/twitter_ideology.

6. Barberá, Pablo., John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?." Psychological Science 26(10):1531-1542.

7. Boxell, Levi., Matthew Gentzkow, and Jesse M. Shapiro. 2017. "Is the Internet Causing Political Polarization? Evidence from Demographics." NBER Working Paper Series No. 23258.

8. Gentzkow, Matthew., and Jesse M. Shapiro. 2006. "Media Bias and Reputation." Journal of Political Economy 114(2): 280-316.

9. Gentzkow, Matthew., and Jesse M. Shapiro. 2011. "Ideological Segregation Online and Offline." The Quarterly Journal of Economics 126(4):1799-1839.

10. Lee, Kevan. 2016. "The Biggest Social Media Science Study: What 4.8 Million Tweets Say About the Best Time to Tweet." Buffer Social. https://blog.bufferapp.com/best-time-to-tweet-research.

11. MediaBias/FactCheck. https://mediabiasfactcheck.com/. First accessed 2017.

12. Mullainathan, Sendhil, and Andrei Shleifer. 2005. "The Market for News." The American Economic Review 95(4):1031-1053.

13. Oremus, Will. 2017. “Twitter's New Order: Inside the changes that could save its business-and reshape civil discourse.” Slate. http://www.slate.com/articles/technology/cover_story/2017/03/twitter_s_timeline_algorithm_and_its_effect_on_us_explained.html.

14. Poole, Keith T., and Howard L. Rosenthal. 2007. “Ideology and Congress.” 2nd. Edition. Transaction Publishers.

15. Punditfact. http://www.politifact.com/punditfact/. Accessed 2017.

16. Rysman, Marc. 2009. “The Economics of Two-Sided Markets.” Journal of Economic Perspectives 23(3):125-143.

17. Shearer, Elisa, and Jeffrey Gottfried. 2017 “News Use Across Social Media Platforms.” Pew Research Center. http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/#many-social-media-news-consumers-still-get-news-from-more-traditional-platforms.

18. Twitter Blog. https://blog.twitter.com/. Accessed 2018.

19. Twitter Developer. https://developer.twitter.com/en.html. Accessed 2017.

20. Twitter Help Center. https://help.twitter.com/en. Accessed 2018.

21. University Libraries: University of Washington. “Savvy Info Consumers: Fake News.” http://guides.lib.uw.edu/research/evaluate/fakenews. Accessed 2017.

22. University of Minnesota Libraries. “Fake News -How to Get to the Bottom of the Story.” https://www.lib.umn.edu/libdata/page.phtml?page_id=5401. Accessed 2017.

23. University of Pittsburgh: University Library System. “Newspapers @ Pitt: Determining Bias and Credibility.” http://pitt.libguides.com/newspapers/bias. Accessed 2017.

APPENDIX C: CHAPTER II TABLES

Table 2.1: News Sources, Reliability, Ideology, and Number of Followers

Screen Name	Reliability	News Ideology	Followers
ABC	2	-0.539	13,631,993
ABCPolitics	2	-0.539	685,827
BBCBreaking	3	-0.597	37,283,877
BBCWorld	3	-0.665	22,822,164
BBCNorthAmerica	3	-0.665	261,714
ACLU	2	-1.164	1,441,517
AlterNet	1	-1.472	140,035
AP	3	-0.496	12,477,917
AssociatedPress	3	-0.496	104,918
blackvoices	2	-0.855	395,459
BreitbartNews	1	1.471	922,991
CBSNews	2	-0.552	6,452,856
CBSPolitics	2	-0.552	276,196
CNBC	2	-0.359	2,920,561
CNBCPolitics	2	-0.359	12,956
CNN	1	-0.612	39,673,601
cnnbrk	1	-0.576	54,340,687
CNNPolitics	1	-0.612	2,625,230
dailykos	1	-1.360	274,178
TheEconomist	2	-0.480	23,056,704
EconUS	2	-0.480	83,541
FoxNews	1	0.760	17,369,783
foxheadlines	1	0.760	147,602
foxnewspolitics	1	1.122	1,989,069
FT	2	-0.351	3,206,968
Heritage	1	1.166	630,670
HuffPost	2	-0.855	11,444,003
HuffPostPol	2	-0.944	1,448,944
latimes	2	-0.736	3,216,103
MoveOn	1	-1.541	293,170
MSNBC	1	-1.087	2,040,523
NRO	2	1.169	283,491
newsbusters	1	1.371	171,519
Newsweek	2	-0.707	3,422,624
NewYorker	2	-0.920	8,503,655
NPR	2	-1.196	7,480,392
nprpolitics	2	-0.918	2,940,793

Table 2.1 (cont.)

Screen Name	Reliability	News Ideology	Followers
nytimes	2	-0.775	41,360,627
NewsHour	2	-0.867	993,013
politico	2	-0.409	3,570,729
Reuters	3	-0.453	19,463,154
ReutersPolitics	3	-0.453	250,951
RollingStone	2	-1.026	6,465,833
Salon	2	-1.171	1,005,701
Slate	2	-0.976	1,797,017
TheAtlantic	2	-0.949	1,663,595
thenation	2	-1.197	1,253,784
thinkprogress	1	-1.352	852,441
TIME	2	-0.719	15,367,498
TODAYshow	2	-0.654	4,397,536
townhallcom	2	1.424	115,749
USATODAY	2	-0.439	3,599,420
usatodayDC	2	-0.439	215,381
washingtonpost	2	-0.566	12,219,550
WSJ	2	-0.066	15,608,598

Note: Number of Followers as of March 11, 2018

Table 2.2: Liberal and Conservative Accounts

User Ideology	Number of Users	Percent
Liberal (ideology ≤ 0)	203	61.33
Conservative (ideology > 0)	128	38.67
Total	331	100.00

Table 2.3: Consumption of Tweets

Consumption	Frequency	Percent
0	41,244	99.30
1	292	0.70
Total	41,536	100.00

Table 2.4: Consumption of Tweets for Liberals and Conservatives

	Liberal		Conservative	
Consumption	Frequency	Percent	Frequency	Percent
0	27,146	99.31	14,098	99.27
1	188	0.69	104	0.73
Total	27,334	100	14,202	100

Table 2.5: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max	Obs
News Reliability	1.873	0.610	1	3	55
News Ideology	-0.472	0.750	-1.541	1.471	55
User Ideology	-0.348	1.305	-2.416	2.410	331

Table 2.6: Number of Accounts, Consumption, and Potential Consumption by News Reliability

Reliability	Number of Accounts	Consumption	Potential Consumption	Consumption Rate
Very High	7	49	7,356	0.67%
High	34	140	23,785	0.59%
Mixed	14	103	10,395	0.99%

Table 2.7: Consumption and Potential Consumption by News Reliability for Liberals

Reliability	Consumption	Potential Consumption	Consumption Rate
Very High	31	4,667	0.66%
High	104	17,733	0.59%
Mixed	53	4,934	1.07%

Table 2.8: Consumption and Potential Consumption by News Reliability for Conservatives

Reliability	Consumption	Potential Consumption	Consumption Rate
Very High	18	2,689	0.67%
High	36	6,052	0.60%
Mixed	50	5,461	0.92%

Table 2.9: Liberal and Conservative News

News Ideology	Number of Sources	Percent
Liberal (ideology ≤ 0)	47	85.45
Conservative (ideology > 0)	8	14.55

Table 2.10: Consumption and Potential Consumption by News Ideology for Full Sample

News Ideology	Consumption	Potential Consumption	Consumption Rate
Liberal	244	37,445	0.65%
Conservative	48	4,091	1.17%

Table 2.11: Consumption and Potential Consumption by News Ideology for Liberals

News Ideology	Consumption	Potential Consumption	Consumption Rate
Liberal	172	26,755	0.64%
Conservative	16	579	2.76%

Table 2.12: Consumption and Potential Consumption by News Ideology for Conservatives

News Ideology	Consumption	Potential Consumption	Consumption Rate
Liberal	72	10,690	0.67%
Conservative	32	3,512	0.91%

Table 2.13: Marginal Effects of Random Effects Probit Estimation of Model (2) for Liberals at Mean Values

Mean $(a_{jt} - b_i)Right_{jt} = 0.529$

Mean $(a_{jt} - b_i)Left_{jt} = -0.086$

Mean $R_{jt} = 1.99$

Varibles	Marginal Effect	95% Lower	95% Upper
Ideological Distance Right	0.009***	0.005	0.013
	(0.002)		
Ideological Distance Left	0.002	-0.009	0.014
	(0.006)		
Reliability	-0.002**	-0.005	-0.889×10^{-4}
	(0.001)		
ρ	0.380		
$ln(\sigma_\mu^2)$	-0.489		
	(0.218)		

Likelihood-Ratio Test of $\rho = 0$: $\bar{\chi}^2 = 212.03$, Prob $>= \bar{\chi}^2 = 0.000$

*p<0.1, **p<0.05, ***p<0.01

Standard errors are in parentheses.

Table 2.14: Marginal Effects of Random Effects Probit Estimation of Model (2) for Conservatives at Mean Values

Mean $(a_{jt} - b_i)Right_{jt} = 0.004$

Mean $(a_{jt} - b_i)Left_{jt} = -1.462$

Mean $R_{jt} = 1.805$

Varibles	Marginal Effect	95% Lower	95% Upper
Ideological Distance Right	-0.032	-0.108	0.045
	(0.039)		
Ideological Distance Left	0.248×10^{-3}	-0.002	0.002
	(0.001)		
Reliability	-0.001	-0.003	0.001
	(0.001)		
ρ	0.093		
$ln(\sigma_{\mu}^{2})$	-2.279		
	(0.524)		

Likelihood-Ratio Test of $\rho = 0$: $\bar{\chi}^2 = 25.38$, Prob $>= \bar{\chi}^2 = 0.000$

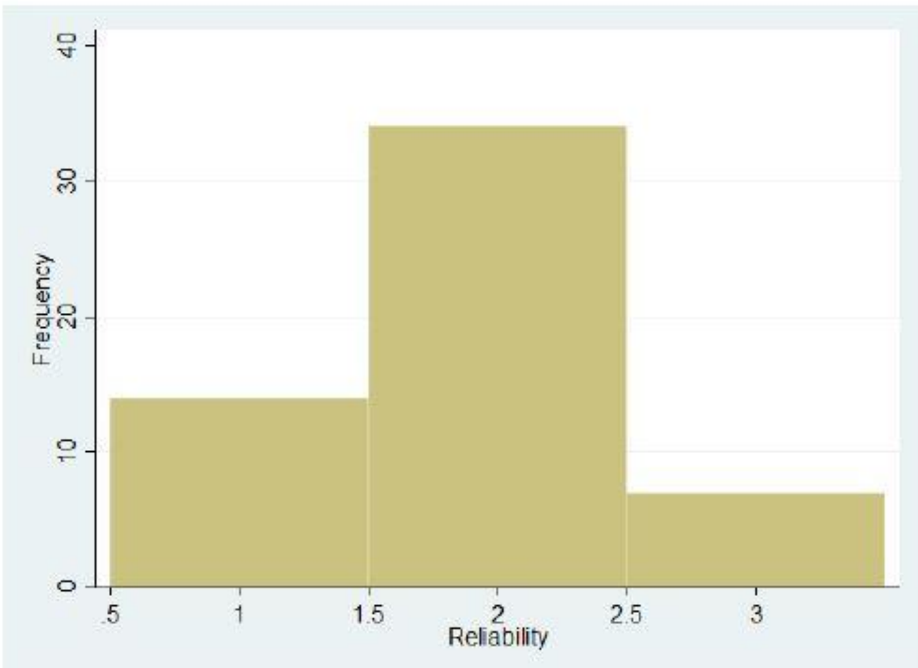
*p<0.1, **p<0.05, ***p<0.01

Standard errors are in parentheses.

APPENDIX D: CHAPTER II FIGURES

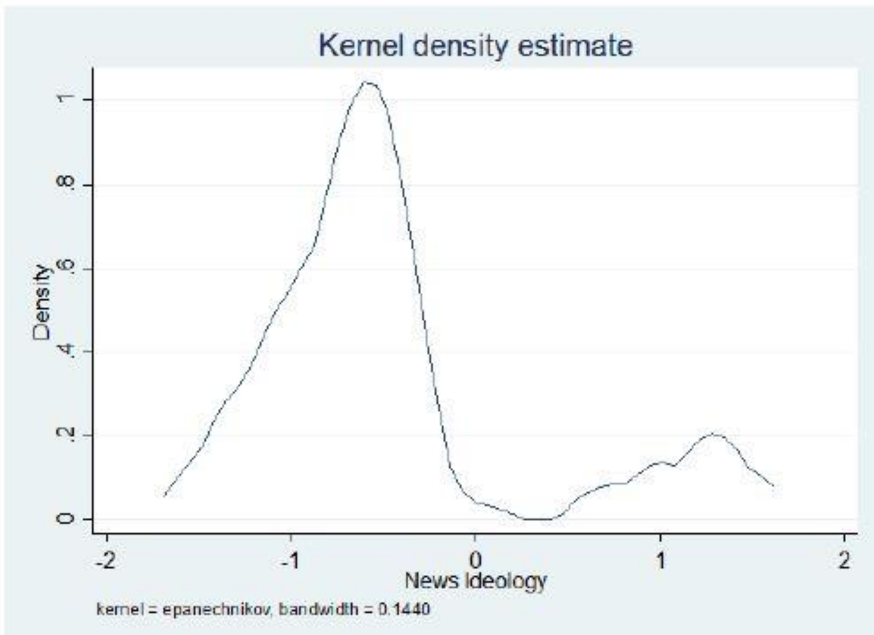
From @CNN Twitter account: The embedded video gives all the information supplied in the link.

Figure 2.1: Consumption without Click (CNN)



Note the histogram is for all news Twitter accounts and not just all news parent organizations.

Figure 2.2: News Reliability Frequency



Note the kernel density is of all news Twitter accounts and not just news parent organizations.

Figure 2.3: Density of News Ideology

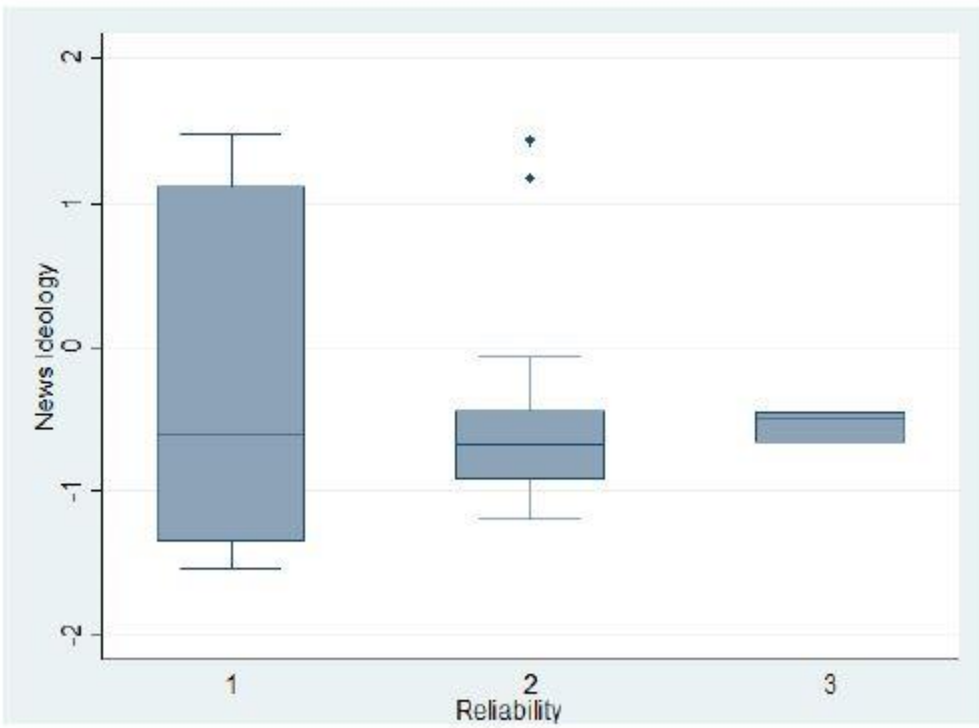


Figure 2.4: Box Plot of News Ideology and News Reliability

CHAPTER III
INSIGHTS INTO THE TWITTER PLATFORM

1 Introduction

Twitter (Twitter 2018) is a social media platform with 330 million active monthly users and 500 million tweets sent per day as of January 2018 (Aslam 2018). As such, a general understanding of the platform is important as it becomes increasingly prevalent in society. The purpose of this study is to describe basic aspects of the Twitter platform. This includes Twitter accounts' networks, types of Twitter accounts, and the type of activity which takes place on Twitter.

I look at active US Twitter accounts. I examine the network of the average account and find accounts have audiences of about 6,000 accounts and receive messages from about 1,500 accounts. This suggests active US accounts have fairly large audiences and potentially receive a large amount of information on a regular basis. I study the percentage of Twitter accounts that are connected to high profile individuals and organizations and how their networks and the frequency of their activity differs from the lower profile accounts. According to the data, accounts associated with high profile individuals and organizations constitute about 1.5% of the total number of accounts. I find these accounts have much larger audiences and receive messages from a slightly larger number of accounts relative to the lower profile accounts. Additionally, I do not find evidence that high profile accounts tweet at a different rate than other Twitter accounts.

Next, I look at the portion of Twitter activity that is connected to the news accounts used in Deaton (2018). I find about 0.5% of activity on Twitter is connected to these news accounts.

For reasons addressed below I believe this is the low end of actual activity on Twitter that relates to the news. These findings suggest the activity I study in Deaton (2018) represents a relatively low fraction of the total activity of active US Twitter accounts.

Finally, I examine the percentage of activity on Twitter constituting original messages relative to activity which is reacting to other messages and examine how this varies based on if the account is high profile or not. I find that the majority of activity on Twitter is a reaction to previous information. This holds true for both high and lower profile accounts, however, high profile accounts have a higher portion of original messages relative to other accounts. In fact, I find that high profile accounts are 13.2 percentage points more likely to post an original message than lower profile accounts. A possible explanation for this finding is high profile accounts have a higher preference for promoting themselves through original messages relative to other accounts. However, without additional information this explanation cannot be validated.

Section 2 gives a brief description of Twitter and how it operates. Section 3 describes the data. Section 4 examines active US Twitter accounts. Section 5 examines activity on Twitter and Section 6 concludes.

2 Twitter

Twitter is a social media platform that allows users to create accounts and post and view short messages called tweets. When a user logs into their account they are directed to their home timeline. The home timeline is populated by tweets from friends, which are accounts an individual has decided to follow. Tweets can be sorted into several categories. A standard tweet is an original message posted by an account. A retweet is a reposting of another account's tweet verbatim. A quote tweet is the same as a retweet except that a small message is added to the original tweet by the reposting account. An additional type of tweet which may populate an account's home timeline is a reply. A reply occurs when one account directly responds to another account's tweet. Replies will enter a particular account's home timeline only if the host account of the home timeline is friends with both the account responsible for posting the original tweet and the account posting the reply. Other content Twitter believes is relevant to an account may be posted to the top of their home timeline. However, the rest of the activity which populates a home timeline consists of tweets from friends posted in reverse chronological order (Oremus 2017).[1] Throughout this study, I refer to accounts an individual chooses to follow as friends and accounts that choose to follow an individual as followers.[2]

3 Data

The data for this study comes from Twitter's Application Programming Interface (API) service (Twitter Developer 2017). From June 20 to June 22, 2017, I collect a random sample of approximately 1% of all tweets sent worldwide between 11 a.m and 3 p.m. Central Time.[3] The data gives information about the content of the tweet, the type of tweet, the account that generated the tweet, and any accounts the tweet references, such as an account that was retweeted.

[1] This is the most up to date information I could find about the Twitter home timeline. For a more detailed discussion on the topic of Twitter home timelines see Oremus (2017).
[2] For additional information on Twitter see the Twitter Help Center (2018) and the Twitter Blog (2018).
[3] The specified time period coincides with noon to 1 p.m. for each of the four time zones in the continental US. This is the peak time for Twitter activity in each time zone (Lee 2016).

I drop all ‘deleted’ tweets and use a simple filter suggested in Barberá et al. (2015) intended to remove inactive and non-US accounts.[4] After applying the filter I am left with 696,907 tweets from 538,050 unique accounts.

The news source accounts used in this study are identical to the ones used in Deaton (2018). It is derived from a list of news sources in Gentzkow and Shapiro (2011) matched with data from MediaBias/FactCheck.com (2017) and estimates from Barberá et al. (2015).[5]
The 55 accounts are presented in Table 3.1

Because of the timing of data collection, it must be kept in mind when examining the data all results are specific to US Twitter accounts active during the middle of a weekday and activity from the same time period. Extrapolating beyond these accounts or timeframe is not possible.

4 Twitter Accounts

First, I examine characteristics of accounts' networks, specifically, the average number of followers and average number of friends for accounts in my sample. This information is presented in Table 3.2. On average, tweets from accounts in my sample are sent to almost 6,000 accounts. This suggests a fairly large audience size for tweets from accounts in my sample. It should be kept in mind the audience size varies greatly as evidenced by the very large standard deviation in the number of followers. Additionally, these statistics do not give evidence of the

[4] The filter drops all accounts who do not identify English as their primary language, who have tweeted less than 100 times, who follow less than 100 accounts, and who are followed by less than 25 accounts.
[5] Please see Deaton (2018) for a more detailed discussion of how these sources were cleaned and collected.

engagement of followers. If a number of followers rarely check their home timeline, the audience for an account's tweets are diminished.

Accounts in my sample are exposed to tweets from approximately 1,500 accounts suggesting a potentially crowded home timeline. Similarly to followers, there is a large standard deviation in the number of friends of an account. The data does not give an indication of the activity of an account's friends. This leaves the possibility of a sparse home timeline if a minority of friends are active.

Next, I study the makeup of different types of accounts on Twitter. Two groups of potential interest on Twitter are verified and non-verified accounts. Twitter defines verified accounts as “accounts of public interest” which can include “accounts maintained by users in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas” (Twitter Help Center 2018). I consider verified accounts to be high profile accounts and non-verified accounts to be lower profile accounts. Table 3.3 breaks down the number of verified and non-verified accounts. It is immediately clear that verified accounts are in the minority making up less than 1.5% of the total number of accounts.

Table 3.4 breaks down the number of followers and friends for verified accounts and Table 3.5 does the same for non-verified accounts. An observation that stands out is the dramatic difference in the average number of followers between verified and non-verified accounts. The ratio of the average number of followers for verified to non-verified accounts is 64.03 to 1. This is evidence high profile accounts have larger potential audiences than lower profile accounts. Because of the large standard deviations in the number of followers for both verified and non-verified accounts, it must again be noted the audience size for individual accounts varies greatly.

Verified accounts also have more friends than non-verified accounts. However, the ratio of the average number of friends for verified accounts to non-verified accounts, 3.62 to 1, is much smaller than for followers. Again, while this may suggest a more crowded home timeline for verified accounts than for non-verified accounts, without further information it is impossible to verify.

Both verified and non-verified accounts have fewer friends than followers. One potential explanation for this is the sample is composed of active accounts. It may be the case that active accounts are followed by large numbers of inactive accounts who simply read tweets instead of tweeting themselves. As a result of the low level of activity, active accounts follow inactive accounts at a lower rate.

5 Tweets

Next, I examine several aspects of tweeting activity on Twitter. Table 3.6 gives the number of tweets by verified accounts and non-verified accounts in the sample. It appears the percentage of tweets from verified accounts and from non-verified accounts is roughly equal to the percentage of verified and non-verified accounts in the sample. This suggests the two groups are similarly active on Twitter.

Another area of interest is the amount of activity on Twitter related to the news. I look into this by looking at the percentage of tweets connected to one of the 55 news accounts in Table 3.1. This includes tweets directly from a news account, which I call original tweets, and tweets responding to a tweet from a news account, which I call reaction tweets. Table 3.7 breaks down what I consider original tweets and what I consider reaction tweets. Table 3.8 gives the percentage of tweets connected to a news account and those that are not connected to a news account. The amount of tweets connected to a news account is less than 1% of the total activity on Twitter. This should be considered a low representation of the activity on Twitter related to the news for several reasons. First, the news sources in Table 3.1 are fairly traditional news outlets and do not include many niche news categories. A good example is sport news outlets such as ESPN. Second, the list of news sources in Table 3.1 is not a comprehensive list of traditional news outlets and thus cannot capture the full extent of news activity on Twitter. Third, not all activity on Twitter related to the news must connect back to a news account. For example, accounts may link a news article in a tweet without ever interacting with a news account. Finally, because of the timing of the data collection, the data may not capture peak times of activity associated with news accounts.

Even with these caveats, the information given in Table 3.8 is useful. It shows the activity on Twitter associated with the news accounts used in Deaton (2018) may be low relative to the overall activity on Twitter. However, if activity related to these particular accounts peaks at a different time of day this information may under represent the amount of activity related to the news accounts in Table 3.1.

Table 3.9 breaks down the sample into original and reaction tweets as categorized in Table 3.7. Interestingly, over 70% of activity on Twitter is some sort of reaction to another tweet. The majority of activity is not the posting of a new status but some sort of interaction with other individuals. This paints Twitter as a platform with the potential for a great deal of conversation between individuals. One caveat here is this information does not explain the content of tweets and therefore does not give an indication of the type of interactions which are occurring and whether they are superficial or something deeper.

Tables 3.10 and 3.11 divide the tweets into original and reaction tweets for verified and non-verified accounts respectively. Both verified and non-verified accounts have a higher percentage of activity which is some form of a reaction to another tweet. However, the ratio of original tweets to reaction tweets for verified accounts, 0.69 original tweets per reaction tweet, is higher than for non-verified accounts, 0.38 original tweets per reaction tweet. This suggests verified accounts may be more likely to produce original tweets than non-verified accounts.

To test this idea further I run a simple probit model regressing the effect of being a verified account on the likelihood of sending an original tweet. The model I fit can be expressed as:

$$Pr(OriginalTweet_i = 1) = \Phi\ (\alpha + \beta VerifiedAccount_i), \text{(1)}$$

where i indicates each individual tweet, $OriginalTweet$ is a dummy variable equal to one if the tweet is an original tweet and zero otherwise, α is a constant, $VerifiedAccount$ is a dummy equal to one if the tweet was sent by a verified account and zero otherwise, and β is the coefficient on $VerifiedAccount$. Φ is the cumulative normal distribution.

The marginal effect at the mean is presented in Table 3.12. The marginal effect is statistically significant at the 1% level and suggests a verified account is 13.2 percentage points more likely to send an original tweet relative to non-verified accounts. A possible explanation for this result is verified accounts are managed by individuals concerned with promoting themselves via original content. Recall verified accounts are managed by journalists, individuals in entertainment, business organizations, politicians and similar high profile individuals and organizations. It is not unreasonable to assume these type of individuals and organizations have a desire to promote themselves and are thus more likely to send original tweets.

6 Conclusion

The purpose of this study is to describe properties of Twitter accounts and Twitter activity using a data set collected with Twitter's API service. I specifically examine active US Twitter accounts and find the average account has an audience of roughly 6,000 accounts and receives tweets from about 1,500 accounts. This suggests the average account has a large audience and, potentially, a crowded home timeline. I find that a rather low percentage of active US Twitter accounts are verified accounts, about 1.5%, and these accounts have a much larger audience than non-verified accounts. While verified accounts also receive tweets from a larger number of accounts than non-verified accounts, the difference is not nearly as dramatic.

When looking at Twitter activity, I find that the percentage of tweets from verified accounts is roughly the same as the percentage of verified accounts and the same is true for non-verified accounts. This suggests verified and non-verified accounts send a similar number of tweets per account on average. I find that activity related to the news accounts from Table 3.1 is less than 1% of all activity on Twitter. This suggests the activity on Twitter connected to the news accounts used in Deaton (2018) is perhaps low relative to the overall activity on Twitter. For reasons discussed above this should be considered lower than the total amount of Twitter activity connected to all sources of news. The last part of activity I examine is the fraction of activity composed by original tweets compared to reaction tweets. I find the majority of activity on Twitter is some form of reaction to other content. While this holds true for verified and non-verified accounts, verified accounts are 13.2 percentage points more likely to send an original tweet than non-verified accounts.

The results of this study are specific to active US Twitter accounts. The sample is representative of activity that occurs during the middle of the day on week days. Causal analysis is beyond the scope of this work and caution should be taken when drawing inferences from this study.

REFERENCES

1. Aslam, Salman. 2018. “Twitter by the Numbers: Stats, Demographics & Fun Facts.” Omnicoreagency. https://www.omnicoreagency.com/twitter-statistics/.

2. Barberá, Pablo., John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. “Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?.” Psychological Science 26(10):1531-1542.

3. Deaton, Maxwell. 2018. “News Consumption in a Social Media Environment: Evidence from Twitter.” Working Paper.

4. Gentzkow, Matthew., and Jesse M. Shapiro. 2011. “Ideological Segregation Online and Offline.” The Quarterly Journal of Economics 126(4):1799-1839.

5. Lee, Kevan. 2016. “The Biggest Social Media Science Study: What 4.8 Million Tweets Say About the Best Time to Tweet.” Buffer Social. https://blog.bufferapp.com/best-time-to-tweet-research.

6. MediaBias/FactCheck. https://mediabiasfactcheck.com/. Accessed 2017.

7. Oremus, Will. 2017. “Twitter's New Order: Inside the changes that could save its business-and reshape civil discourse.” Slate. http://www.slate.com/articles/technology/cover_story/2017/03/twitter_s_timeline_algorithm_and_its_effect_on_us_explained.html.

8. Twitter. https://twitter.com/. Accessed 2018.

9. Twitter Blog. https://blog.twitter.com/. Accessed 2018.

10. Twitter Developer. https://developer.twitter.com/en.html. Accessed 2017.

11. Twitter Help Center. https://help.twitter.com/en. Accessed 2018.

APPENDIX E: CHAPTER III TABLES

Table 3.1: News Sources with Number of Followers

News Screen Name	Number of Followers	News Screen Name	Number of Followers
ABC	13,631,993	latimes	3,216,103
ABCPolitics	685,827	MoveOn	293,170
BBCBreaking	37,283,877	MSNBC	2,040,523
BBCWorld	22,822,164	NRO	283,491
BBCNorthAmerica	261,714	newsbusters	171,519
ACLU	1,441,517	Newsweek	3,422,624
AlterNet	140,035	NewYorker	8,503,655
AP	12,477,917	NPR	7,480,392
AssociatedPress	104,918	nprpolitics	2,940,793
blackvoices	395,459	nytimes	41,360,627
BreitbartNews	922,991	NewsHour	993,013
CBSNews	6,452,856	politico	3,570,729
CBSPolitics	276,196	Reuters	19,463,154
CNBC	2,920,561	ReutersPolitics	250,951
CNBCPolitics	12,956	RollingStone	6,465,833
CNN	39,673,601	Salon	1,005,701
cnnbrk	54,340,687	Slate	1,797,017
CNNPolitics	2,625,230	TheAtlantic	1,663,595
dailykos	274,178	thenation	1,253,784
TheEconomist	23,056,704	thinkprogress	852,441
EconUS	83,541	TIME	15,367,498
FoxNews	17,369,783	TODAYshow	4,397,536
foxheadlines	147,602	townhallcom	115,749
foxnewspolitics	1,989,069	USATODAY	3,599,420
FT	3,206,968	usatodayDC	215,381
Heritage	630,670	washingtonpost	12,219,550
HuffPost	11,444,003	WSJ	15,608,598
HuffPostPol	1,448,944		

Note: Number of Followers as of March 11, 2018

Table 3.2: Active US Twitter Account Followers and Friends

Variable	Mean	Std. Dev.	Min.	Max.	Obs
Followers	5,926.27	148,567.50	25	68,787,910	538,050
Friends	1,516.77	8,943.27	100	1,270,095	538,050

Table 3.3: Verified and Non-verified Accounts

Variable	Number of Accounts	Percent
Verified	7,634	1.42
Non-Verified	530,416	98.58
Total	538,050	100.00

Table 3.4: Active US Twitter Account Followers and Friends for Verified Accounts

Variable	Mean Number of Followers	Std. Dev.	Min.	Max.	Obs.
Followers	200,309.31	1,204,579.60	170	68,787,910	7,634
Friends	5,293.00	30,773.19	100	1,207,366	7,634

Table 3.5: Active US Twitter Account Followers and Friends for Non-Verified Accounts

Variable	Mean Number of Friends	Std. Dev.	Min.	Max.	Obs.
Followers	3,128.61	30,941.25	25	4,806,948	530,416
Friends	1,462.42	8,203.49	100	1,270,095	530,416

Table 3.6: Number of Tweets by Verified and Non-Verified Accounts

Variable	Number of Tweets	Percent
Verified	9,345	1.34
Non-Verified	687,562	98.66

Table 3.7: Description of Original and Reaction Tweets

Original Tweets	Reaction Tweets
Tweet	Retweet
	Quote Tweet
	Retweet of Quote Tweet
	Reply
	Retweet of Reply
	Quote Tweet of Reply
	Retweet of Quote Tweet of Reply
	Quote Tweet of a Quote Tweet
	Retweet of a Quote Tweet of a Quote Tweet

Table 3.8: Number of Tweets Related to News Source

Variable	Number of Tweets	Percent
From News Source	4,240	0.61
Not From News Source	692,667	99.39

Table 3.9: Number of Original and Reaction Tweets

Variable	Number of Tweets	Percent
Original	193,716	27.80
Reaction	503,191	72.20

Table 3.10: Number of Original and Reaction Tweets for Verified Accounts

Variable	Number of Tweets	Percent
Original	3,818	40.86
Reaction	5,527	59.14

Table 3.11: Number of Original and Reaction Tweets for Non-Verified Accounts

Variable	Number of Tweets	Percent
Original	189,898	27.62
Reaction	497,664	72.38

Table 3.12: Marginal Effects of Probit Regression (1) at Mean

Mean $VerifiedAccount = 0 : 0.99$

Mean $VerifiedAccount = 1 : 0.01$

Variables	Marginal Effects	95% Lower	95% Upper
Verified Account	0.132*** (0.005)	0.122	0.142

*p<0.1, **p<0.05, ***p<0.01

Standard errors are in parentheses.