ABSTRACT

Title of dissertation: ESSAYS IN EMPIRICAL INDUSTRIAL ORGANIZATION

Matthew Chesnes, Doctor of Philosophy, 2009

Dissertation directed by: Professor John Rust
Professor Ginger Jin
Department of Economics

Chapter 1: Capacity and Utilization Choice in the US Oil Refining Industry

This paper presents a new dynamic model of the operating and investment decisions of US oil refiners. The model enables me to predict how shocks to crude oil prices and refinery shutdowns (e.g., in response to hurricanes) affect the price of gasoline, refinery profits, and overall welfare. There have been no new refineries built in the last 32 years, and although existing refineries have expanded their capacity by almost 13% since 1995, the demand for refinery products has grown even faster. As a result, capacity utilization rates are now near their maximum sustainable levels, and when combined with record high crude oil prices, this creates a volatile environment for energy markets. Shocks to the price of crude oil and even minor disruptions to refining capacity can have a large effect on the downstream prices of refined products. Due to the extraordinary dependence by other industries on petroleum products, this can have a large effect on the US economy as a whole.

I use the generalized method of moments to estimate a dynamic model of capacity and utilization choice by oil refiners. Plants make short-run utilization rate choices to maximize their expected discounted profits and may make costly long-term investments in capacity to meet the growing demand and reduce the potential for breaking down. I show that the model fits the data well, in both in-sample and out-of-sample predictive tests, and I use the model to conduct a number of counterfactual experiments. My model predicts that a 20% increase in the price of crude oil is only partially passed on to consumers, resulting in higher gasoline prices, lower profits for the refinery, and a 45% decrease in total welfare. A disruption to refining capacity, such as the one caused by Hurricane Katrina in 2005, raises gasoline prices by almost 16% and has a small negative effect on overall welfare: the higher profits of refineries partially offsets the large reduction in consumer surplus. As the theory predicts, these shocks have a smaller effect on downstream prices when consumer

demand is more elastic, resulting in a larger share of total welfare going to the consumer.

Chapter 2: Consumer Search for Online Drug Information

Consumers are increasingly turning to the internet and using search engines to find information on medicinal drugs. Between 2001 and 2007, the number of adults using the internet as an alternative source of health information doubled. At the same time, online and offline advertising spending by drug companies is growing rapidly. I seek to understand how consumers use search engines to find drug information and how this activity is influenced by direct to consumer advertising.

I utilize a database of user click-through data from America Online to analyze the search behavior of consumers seeking drug information online. Compared with other searches, users submitting drug-related queries are more likely to click on more than one result in a search session, and when they do, they click more rapidly through the results and tend to migrate away from dot-com sites and toward those ending in dot-org and dot-net. Offline advertising on a drug serves to increase the frequency and intensity of these searches.

Chapter 3: Drug Information via Online Search Engines

This paper utilizes a database of organic and sponsored search results from four large search engines to analyze the supply of drug-related information available on the internet. I show that the information varies significantly across search engines, domain extensions, and between organic and sponsored results. Regression results reveal that websites with relatively more promotional content are pushed down in the search results while informational sites (including those ending in dot-gov and dot-org) are more likely to appear on page one of the results.

ESSAYS IN EMPIRICAL INDUSTRIAL ORGANIZATION

by

Matthew William Chesnes

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor John Rust, Co-Chair
Professor Ginger Jin, Co-Chair
Professor Peter Cramton
Professor Pablo D'Erasmo
Professor Erik Lichtenberg

Acknowledgments

I thank my advisors, John Rust, Ginger Jin and Peter Cramton, for their invaluable guidance, as well as Pablo D'Erasmo and Erik Lichtenberg for participating in my defense. I am grateful to the Department of Economics at the University of Maryland and the Economic Club of Washington for their financial support. I also want to thank John Shea, Mark Duggan, Adam Copeland, David Givens, Ariel BenYishay, and seminar participants at the University of Maryland, the Federal Reserve Board of Governors, the International Industrial Organization Conference, and La Pietra-Mondragone Workshop in Economics for their suggestions and comments. All remaining errors are my own.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

API	American Petroleum Institute
DTCA	Direct-To-Consumer Advertising
EIA	Energy Information Administration
FCC	Fluid Catalytic Cracking
FDA	Food and Drug Administration
GMM	Generalized Method of Moments
HHI	Herfindahl-Hirschman Index
HTML	HyperText Markup Language
IANA	Internet Assigned Numbers Authority
MTBE	Methyl Tertiary Butyl Ether
NAMCS	National Ambulatory Medical Care Survey
NDC	National Drug Code
OPEC	Organization of the Petroleum Exporting Countries
OTC	Over-The-Counter
PADD	Petroleum Administration for Defense Districts
RLD	Reference Listed Drug
URL	Uniform Resource Locator

Chapter 1

Capacity and Utilization Choice in the US Oil Refining Industry

1.1 Introduction

The United States is the largest consumer of crude oil in the world and this resource accounts for 40% of the country's total energy needs.[1] Although a majority of this oil comes from foreign sources, almost all is refined domestically. Refineries distill crude oil into a large number of products such as gasoline, distillate (heating oil), and jet fuel. While much attention has been paid to the upstream crude oil production industry (see Hamilton (1983) and Hubbard (1986)), and the downstream retail sector (see Borenstein (1991 & 1997)), very little research has focused on the role of the refining industry. Two important dynamic decisions faced by refiners are their investment in capacity and the utilization rate at which they run their plant. These choices are defined over different time horizons.[2] The optimal choice of capacity accumulation, i.e., the increased ability to distill crude oil into higher valued products, is a long-term decision. Capacity is expensive to build and may take time to come online so forecasts of future market conditions are crucial. A shorter-term problem involves a refiner's choice of capacity utilization. This rate measures the intensity with which a firm uses its capital, which for a refinery may include the use

[1] Source: *2007 Annual Energy Review*, Energy Information Administration (EIA).

[2] In addition, they must solve a complicated linear programming problem because their relative output prices are constantly changing and they have the choice of utilizing different types of crude oil, some of which are better adapted to producing certain products.

of boilers, distillation columns, and downstream cracking units.[3]

The refiner's problem is further complicated by changing market conditions, geopolitical tensions, and unexpected events, such as hurricanes. The largest component of refiners' output is gasoline. New alternative technologies, such as hybrid cars, and changing perceptions on the environmental impact of gas-powered vehicles has affected the sensitivity of consumer demand to the price of gasoline.[4] This affects the ability of refiners to pass through shocks to the price of crude oil resulting from, for example, reduced production from OPEC countries or a war in the Middle East. With about one-half of US refining capacity located along the Gulf of Mexico, the potential for hurricanes can also dramatically affect the ability of the industry to supply a consistent flow of gasoline and other products to the rest of the country.

This paper develops and estimates a new dynamic model of the operating and investment decisions of US oil refiners. These refiners face the possibility of breaking down if they run their plant too intensively, so they make costly investments in capacity to reduce this potential and to meet the growing demand for their products. My model assumes that firms are Cournot competitors in the refined product market. With many small firms, each is approximately a price-taker in the market, so the model of Kreps and Scheinkman (1983), with quantity pre-commitment (capacity choice) and Bertrand price competition, is similar to my approach. The model enables me to predict how shocks to crude oil prices and refinery shutdowns (e.g., in response to hurricanes) affect the price of gasoline, refinery profits, and overall

[3]More details on the refining process can be found in section 2 and in appendix A.

[4]Knittle et al. (2008) and Espey (1996) both study the recent changes in consumers' price elasticity of demand for gasoline.

welfare.[5] I also estimate how a change in the price sensitivity of consumers may affect the results of these shocks, particularly in regards to the division of welfare between the refiner and the consumer.

I estimate a fully dynamic model of the oil refining industry incorporating key decisions made by plants which affect both contemporaneous and future profitability. The refining industry is inherently forward-looking and decisions made today rely heavily on forecasts of future market conditions. A static model would not, for example, account for the increased breakdown potential of a plant from high utilization rates or the appropriate long-term investments of a refiner facing rising crude oil costs and uncertain demand. My estimation algorithm involves classic policy function iteration nested inside a GMM optimization, which allows me to compute the equilibrium value and policy functions.[6] This approach allows me to run various counterfactual experiments and determine the optimal policy and future discounted profits of each firm. Several recent papers, including Bajari et al. (2007) and Ryan (forthcoming), estimate dynamic models of firm behavior using a 2-step method that reduces the computational complexity of finding the structural parameters, but does not allow one to compute the equilibrium under counterfactual environments.

My model predicts that a 20% increase in the price of crude oil is only partially passed on to consumers, resulting in a 13% increase in gasoline prices, lower profits for the refinery, and a 45% decrease in total welfare. The pass-through result is fairly close to the historic rate of about 50%.[7] Consumer surplus falls following the

[5] I define total welfare to be the sum of consumer surplus and refiner profit.

[6] See Rust (2008).

[7] See Borenstein and Shepard (1996) and Goldberg and Hellerstein (2008) for related literature on price pass-through.

shock, but the change in the overall distribution of welfare depends on the sensitivity of consumer demand to the prices of refined products. More sensitive consumers sacrifice less and receive a larger share of the (smaller) surplus. I also show that a disruption to refining capacity, such as the one caused by Hurricane Katrina in 2005, raises gasoline prices by almost 16% and has a small negative effect on overall welfare: the higher profits of operating refineries partially offset the large reduction in consumer surplus. When Hurricane Katrina hit the Gulf Coast in August 2005, the actual wholesale gasoline price rose by 14% the following month.

Much of the literature on retail gasoline markets has focused on the asymmetric response of gasoline prices to crude oil shocks, the so-called *rockets and feathers* phenomenon (for example, see Borenstein (1997), Bacon (1991), and Noel (2007)).[8] Recent research on the wholesale gasoline market includes Hastings et al. (2008), which analyzes wholesale prices and the effects of new environmental regulations, and studies by The Government Accountability Office (2006), the Federal Trade Commission (2006), and the Energy Information Administration (2007).

To my knowledge, this is the first dynamic model of the US oil refining industry. Refiners play an important role as an intermediary between upstream crude suppliers and downstream retail markets. A complete analysis of the oil industry must account for the important effects of the refiners' dynamic decisions. I show that the model fits the data well and can be used to generate insights into the pass-through of crude oil shocks and the impacts of refinery shutdowns on consumers. The model's

[8]The market power gained by the refining industry due to a tight capacity environment is one potential explanation. Others include search costs in the retail market, inventory management by consumers who may fill their tank more frequently as prices rise, but are less eager to "top-off" when prices are falling, and adjustment costs at the refinery.

main features include a dynamic decision process, long-term investment choices, and the possibility of plant break-down. The framework could be applied to other energy markets as well as industries, such as shipping, that make large investments in capacity based on expectations of future market conditions.

The remainder of this paper is organized as follows. In section 2, I provide an overview of the oil refining industry to better understand the complicated problem facing the refiner. I describe my data in section 3 and lay out a dynamic model of the industry in section 4. Section 5 provides the details of my empirical strategy and I summarize the fit and results of the model in section 6. Finally, in section 7, I use my estimated parameters to run several counterfactual experiments involving shocks to the price of crude oil, refining capacity, and consumers' price elasticity of demand. Section 8 concludes and provides a discussion of potential extensions.

1.2 The US Oil Refining Industry

The oil industry is broadly comprised of several vertically oriented segments. They include crude oil exploration and extraction, refineries which distill crude oil into other products, pipeline distribution networks, terminals which store the finished product near major cities, and tanker trucks which transport products to retail outlets.[9] The largest refined product, gasoline, accounts for about 50% of total production, while distillate makes up another quarter. A full 68% of output from the oil refining industry is used in the transportation industry. Figures 1.1

[9] 75% of terminals in the US are owned by companies not involved in the upstream exploration and refining.

and 1.2 provide a description of the production process and average product yields. The main distillation process produces some final products like gasoline, but it is complemented by other units that extract more of the highest valued products. Technical details of the refining process and background on the types of crude oil available can be found in the appendix.

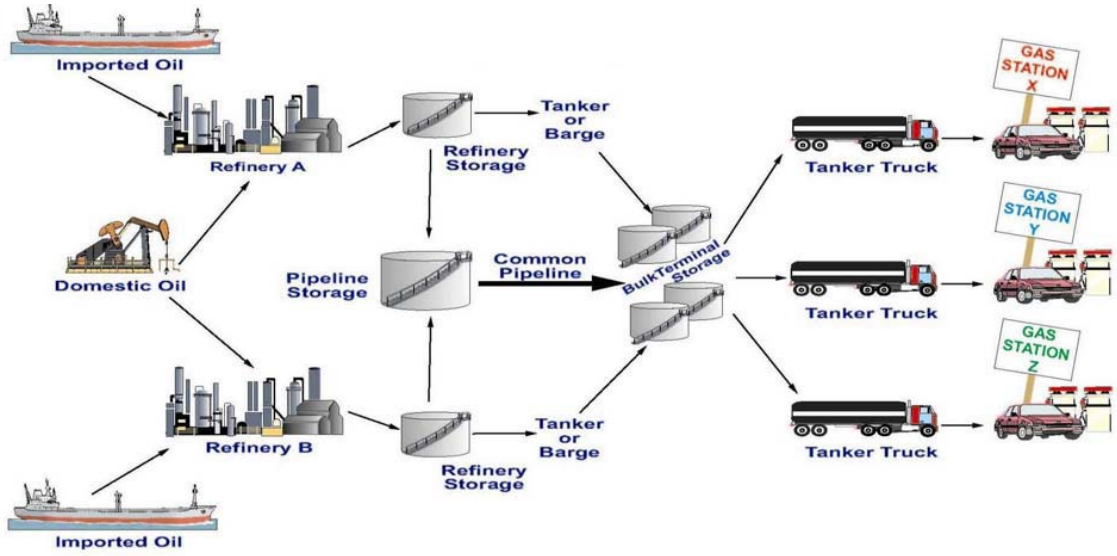


Figure 1.1: Production Process

The market for refined oil products is large and growing, with the US consuming 388 million gallons of gasoline each day and one quarter of the world's crude oil.[10] Aside from refining crude oil into gasoline, refineries produce many products that are important inputs into other industries. Retail gasoline prices have recently experienced increased variability in the US and in summer 2008 hit an all time high of $4.11 per gallon. Wholesale prices peaked around $3.40 a gallon in the same period.[11] Many justify the high prices as a result of the growing demand for gasoline

[10]Annual world consumption of crude oil totals 30 billion barrels, of which 7.5 billion barrels comes from the US. About 60% of crude oil used by refineries is imported and US consumption of refined gasoline represents 40% of world consumption.

[11]US regular gasoline, source: EIA.

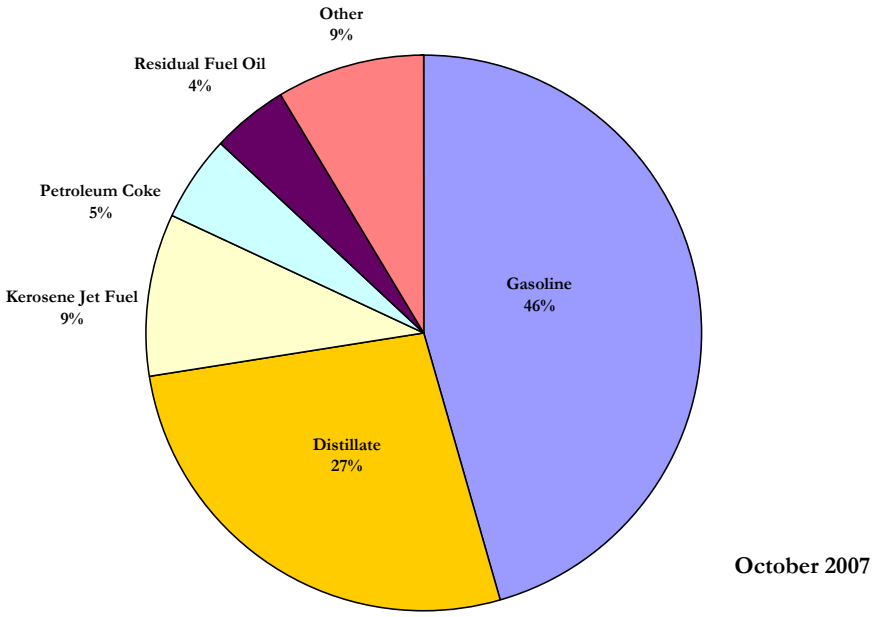


Figure 1.2: Average Yields

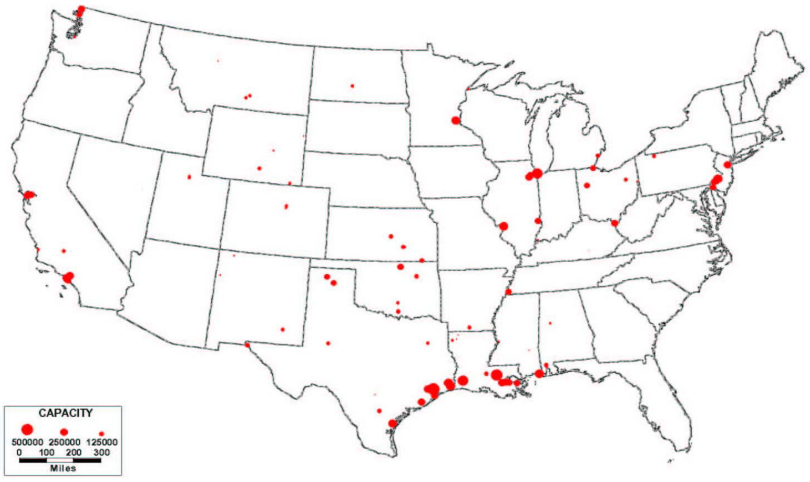
and supply limitations, including the scarcity of crude oil, Middle East uncertainty, hurricanes, and the OPEC cartel. Others claim the high prices result from coordinated anticompetitive behavior by big oil companies. It may be that the strategic capacity investment and utilization choices by oil refineries play a significant role in affecting downstream prices, profits, and consumer welfare.

1.2.1 Competition

Concentration

The refining industry is fairly competitive, with 144 refineries owned by 54 refining companies in January 2006. About one-half of US production occurs near the Gulf of Mexico in Texas and Louisiana, though there are significant operations in the Northeast, the Midwest, and California. During World War II, the country was

divided into Petroleum Administration for Defense Districts (PADDs) to aid in the allocation of petroleum products. Figure 1.3 displays a map of refinery locations along with delineations of PADDs and PADD districts. PADDs are often used by regulators such as antitrust authorities when assessing market concentration. See table A.2 in appendix D for concentration ratios and Herfindahl-Hirschman Indices (HHIs) for various PADDs and regions at the refiner level. The degree of market concentration is clearly dependent upon how one defines the relevant geographic market.[12]



PADD	District	Area
I	1	East Coast
II	2	Midwest
	3	Upper Midwest
	4	Central Plains
III	5	Louisiana
	6	Texas
IV	7	New Mexico
	8	Rockies
V	9	West Coast

Figure 1.3: Refinery Locations (Scaled by Capacity)

Market Definition

While retail markets for gasoline tend to be very small, markets for wholesale gasoline are relatively large due to the extensive pipeline network use to transport most refined products. While a PADD may have roughly approximated a market in 1945, these delineations were made before the pipeline network had been fully developed,

[12] At the national level, the top four refiners (who each own multiple refineries) controlled 44.1% of the market in 2007. The HHI for refiners on the Gulf Coast was about 1,100, which would be classified as *moderately concentrated* according to the Horizontal Merger Guidelines.

so they are now just a convenient way to report statistics on the industry.[13] A map of major crude oil and production piplines is shown in figure 1.4. With important pipelines connecting the Gulf Coast production center to the population centers in the Northeast and the Midwest, I combine PADDs 1, 2, and 3 into one large market for wholesale gasoline. I denote the Rocky Mountain region, PADD 4, as another market, because it is isolated from the rest of the country and imports only limited refined product from other regions. Finally, my third market is the West Coast, PADD 5, which includes California, a state that, due to strict environmental regulations, is limited in its ability to use products that are refined in other states.

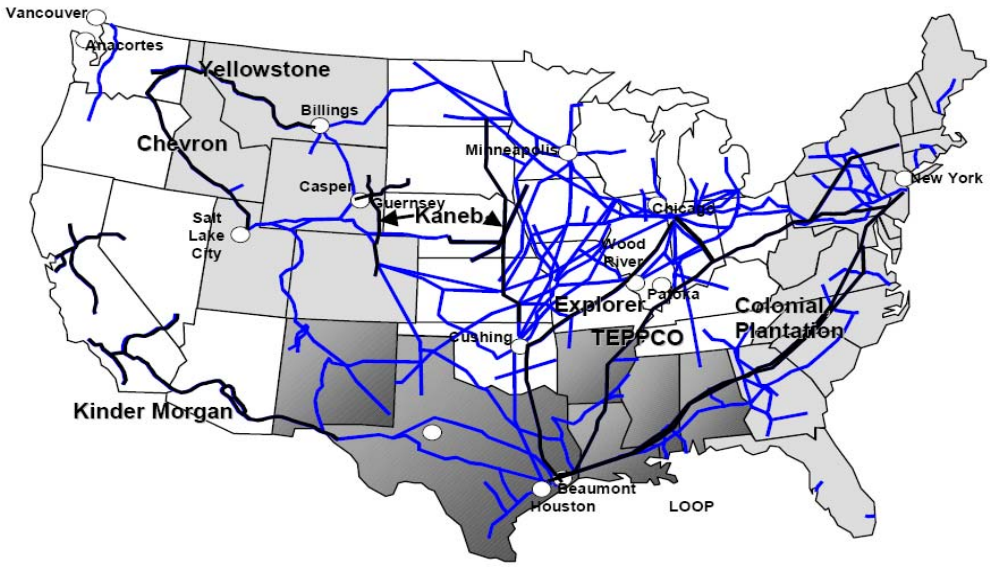


Figure 1.4: Major Refined Product Pipelines

Aside from the domestic refining industry, US refiners face limited competition from abroad. While the US is very dependent on foreign oil, domestic production accounts for about 90% of US gasoline consumption, though the import share has

[13]For instance, the Colonial pipeline, which runs from the Gulf Coast up to the Northeast, was built in 1968. Pipelines now carry 70% of all refined products shipped between PADDs.

grown since the mid 1990s. These imports come primarily into the Northeast, which receives 45% of its supply from sources, such as the US Virgin Islands, the United Kingdom, the Netherlands, and Canada. Recent US regulations limiting certain types of fuel additives combined with increased European dependence on diesel fuel has limited the ability of US markets to rely on foreign imports.

1.2.2 Capacity and Utilization

Capacity utilization rates at US refineries have been steadily rising and are now at their maximum sustainable levels. From 2000 to 2008, the average utilization rate in US manufacturing industries was 77%, while in the refining industry it was 91%.[14] At the same time, no new refineries have been built in the US since 1976. In fact, many plants have closed and the number of refineries has fallen from 223 in 1985 to just 144 today. However, most of these closures were small and inefficient plants, and those that remain have expanded, so total operable capacity has grown from 15.6 million barrels per day (bbl/day) in 1985 to almost 17 million bbl/day today. However, this figure is lower than in 1981, when capacity was 18.6 million bbl/day. The overall number of refineries along with their production capacity are displayed in figure 1.5. The average plant size has increased from 74,000 bbl/day in 1985 to almost 124,000 bbl/day in 2007.

Building a new refinery is very expensive, and environmental requirements and permits create significant hurdles.[15] Evidence from a 2002 US Senate hearing

[14] See `http://www.federalreserve.gov/releases/G17/caputl.htm`.

[15] One of the few new plants in development is in Yuma, Arizona. The builder of the 150,000 bbl/day refinery has spent 30 million dollars over 6 years to acquire all the permits. If not blocked,

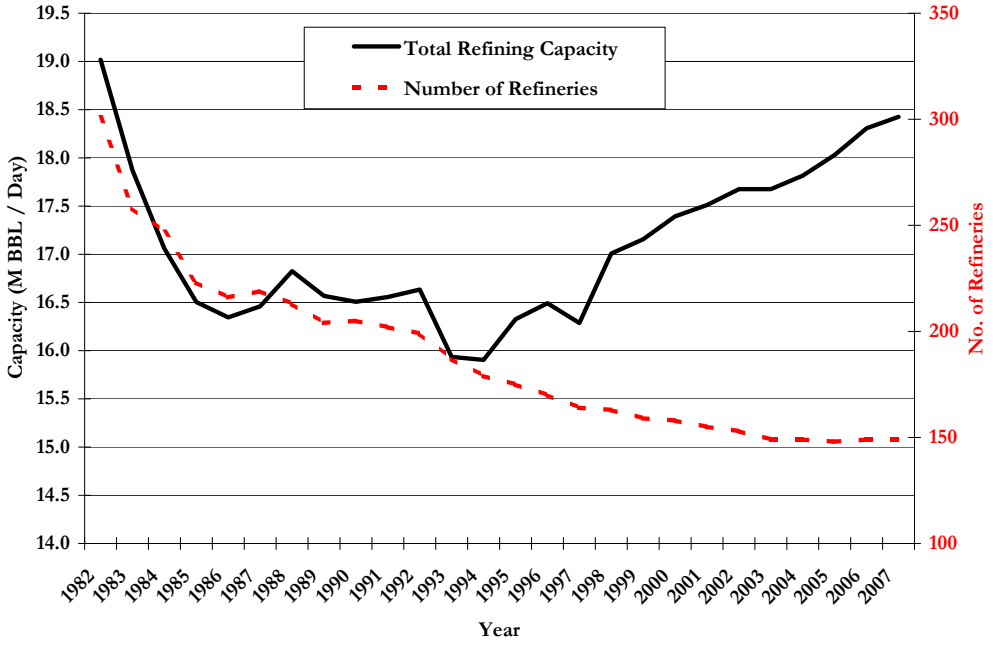


Figure 1.5: Capacity and Number of Refineries

estimated the cost of building a 250,000 bbl/day refinery at around 2.5 billion dollars, with a completion time of 5-7 years (Senate (2002)). This assumes the various environmental hurdles and community objections are satisfied. No one wants a dirty refinery operating near them.[16] In May 2007, the chief economist at Tesoro, Bruce Smith, was quoted as saying that the investment costs in building a new refinery are so high that "you'd need 10 to 15 years of today's margins [at the time, around 20%] to pay it back."[17] Even without new refineries, existing refineries have invested to expand capacity. The distribution of historical investment rates is shown in figure 1.6. While the mean investment has been 1.3% per year, the median is zero

construction on the new refinery will begin in 2009.

[16]Commonly referred to as "NIMBY," an acronym for Not In My Back Yard.

[17]The National Petrochemical & Refiners Association estimates that the average return on investment in the refining industy between 1993-2002 was 5.5%. The S&P 500 averaged over 12% for the same period. See "Lack of Capacity Fuels Oil Refining Profits" available online at `http://www.npr.org/templates/story/story.php?storyId=10554471` (downloaded: 09/13/2008).

as plants tend to make very infrequent investments. Even restricting the sample to non-zero changes as shown in the graph, investments tend to be small, with almost 85% of the non-zero changes less than 10%.

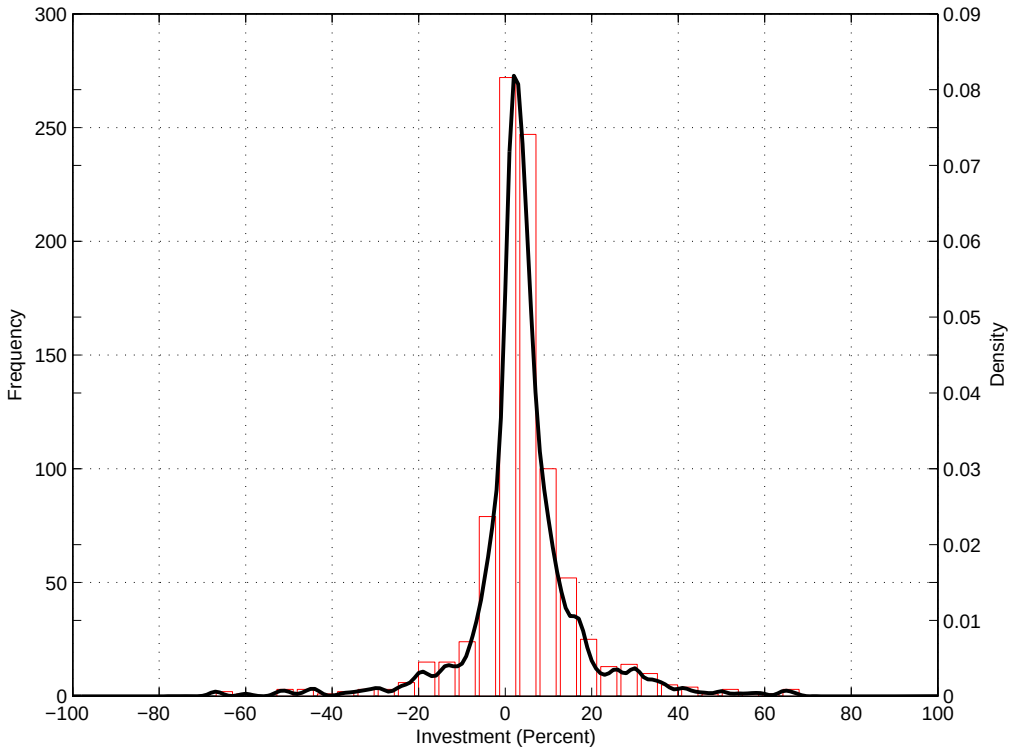


Figure 1.6: Non-Zero Changes in Capacity, All Plants, 1986-2007

Although oil refining has historically been an industry plagued by thin profit margins, oil producers are now starting to make higher profits from their refining business. One simple measure of the profit margin at a refinery is the "crack spread." For every barrel of crude oil the refinery uses, technological constraints require that about half of it goes into gasoline production and about a quarter into distillate. So the crack spread, expressed in dollars per barrel, is calculated as:

$$Crack = \frac{1 * Price(distillate) + 2 * Price(gasoline) - 3 * Price(crude\ oil)}{3} \quad (1.1)$$

The crack spread along with the utilization rates of refineries are shown in figure 1.7. The crack spread hit a record high of nearly $30 per barrel in July 2006. Some argue that based on this measure of profitability, it is surprising that more refiners have not overcome the setup costs and entered this industry. The increase in the crack spread after 2000 occurred after the utilization rate had already been at a very high level. This may imply that a refiner's ability to pass through their crude oil cost has changed since 2000, perhaps due to the scarcity of crude oil, an increase in industry concentration, or an increase in the demand for gasoline.

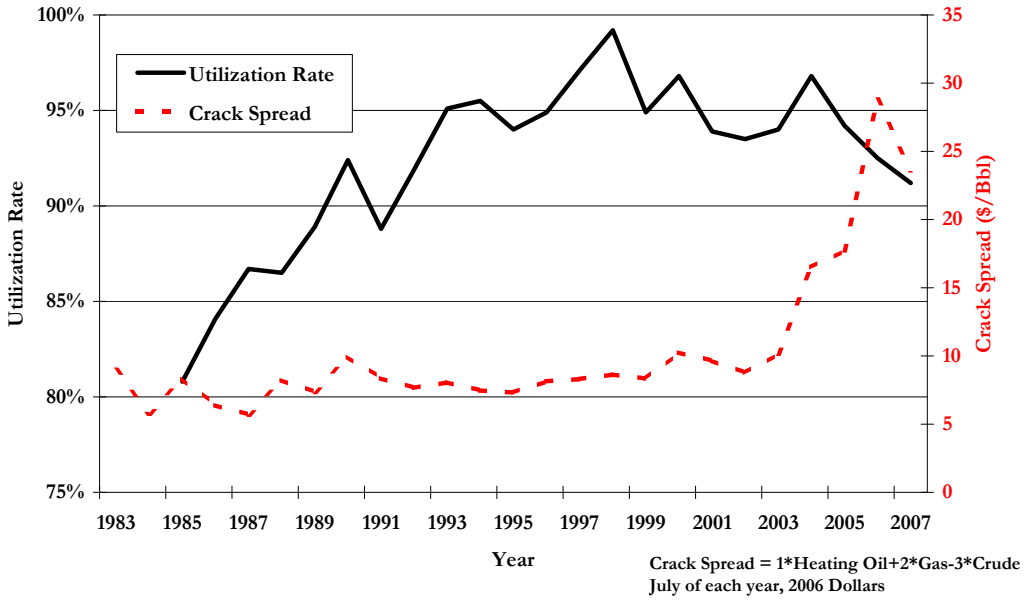


Figure 1.7: Capacity Utilization Rate and Crack Spread

While total refining capacity has risen in the past 10 years, it has not kept up with demand growth. Capacity of oil refiners has increased by 10% in the past 10 years, while demand for gasoline has increased about 17%. The gap has been filled by higher utilization rates and, to a lesser degree, growing imports. New

regulations requiring the shift from MTBE[18] oxygenates to ethanol poses a problem for this segment of supply because foreign refiners have not invested in the facilities to produce ethanol blended gasoline. With capacity tight and supply alternatives limited, even a minor supply disruption (or a major one like Hurricane Katrina) can have a large price impact.[19]

1.2.3 Refinery Maintenance and Outages

An oil refinery is a complex operation that requires frequent maintenance, ranging from small repairs to major overhauls.[20] The regular maintenance episodes tend to be short and have minimal impact on production as they are strategically scheduled for low demand periods. Unplanned major outages, by definition, can take place at any time and can have a major impact on production capability. The EIA divides refinery outages into four classes, summarized in table 1.1.

Table 1.1: Refinery Downtime

Type	Typical Length of Outage	Frequency
Planned Shutdowns	1-2 Weeks	Every year
Unplanned Shutdowns	2-4 Weeks	-
Planned Turnarounds	3-9 Weeks	Every 3-5 years
Emergency Shutdowns	Varies	-

Source: EIA.

[18]Methyl Tertiary Butyl Ether.

[19]Following Hurricane Katrina on 9/23/05, capacity fell by 5 MBbl/Day. This represented a full one third of US refining capacity. Inventories are also limited as there is only about 20-25 days worth of gasoline in storage at any time.

[20]Refinery maintenance is crucial not only for production sustainability, but also for the safety of the plant. A 2005 fire at BP's Texas City refinery killed 15 workers and injured over 100 more.

Planned turnarounds are major refinery overhauls, while planned shutdowns bridge the gap between turnarounds. Unplanned shutdowns involve unexpected issues that may allow for some strategic planning of the downtime, but often may force a refinery to reduce production sub-optimally. Finally, emergency shutdowns are those that cause an immediate plant breakdown like a refinery fire.

Organization for planned turnarounds typically start years in advance, and cost millions of dollars to implement, in addition to the revenue lost from suspending production. Due to the hiring of outside personnel, major refineries often have to plan these turnarounds at different times because of the shortage of skilled labor to implement them. Given the typical seasonal variation in product demand, the ideal periods for maintenance are the first and third quarter of the year, though in some northern refineries, cold winter weather forces shifts in planned downtimes.

Even though refineries consist of several components, such as distillation columns, reformers and cracking units, these components are dependent on one another so a breakdown of any one component can affect the production capability of the entire refinery. Downstream units include hydrocrackers, reformers, fluid catalytic cracking (FCC) units, alkylation units, and coking units. They are responsible for breaking down hydrocarbons into more valuable products and removing impurities such as sulfur. For example, in a typical refinery, only 5% of gasoline is produced from the primary distillation process; the rest comes hydrocrackers (5%), reformers (30%), FCC and alkylation units (50%), and coking units (10%). Not all refineries have all of these components, so such refineries are even more affected when one component goes down (EIA (2007)).

At the PADD level, EIA reports that in the 1999-2005 period, refineries experienced reductions in monthly gasoline and distillate production of up to 35% due to outages. At the monthly frequency, there is little effect of outages on product prices. This is primarily because most (planned) outages occur during the low-demand months when markets are not tight; most outages last less than a month; and the availability of imports, increased production from other refineries, and inventories provide a cushion to supply. However, major outages, like those caused by a hurricane, still affect the downstream prices and profitability of all refineries.

Overall, the oil refining industry features several economic *puzzles*, some of which I explore in this paper. While the industry is relatively competitive, refiners have recently been earning significant profits, as measured by the growing crack-spread. However, entrants have yet to overcome the regulations and costs of setting up a new plant and existing firms have been cautious in their expansion. As a result, plants run at high rates of utilization, which leads to instability in the face of unexpected capacity disruptions.

1.3 Data

The EIA publishes data on the oil refining industry at various frequencies and levels of aggregation.[21] I observe monthly district level data, which is publicly

[21] Although monthly plant level data is collected from individual refineries on EIA form 810, this data remains proprietary and unavailable to academic researchers. A new program, joint with the National Institute for Statistical Sciences (NISS), called the NISS-EIA Energy Micro Data Research Program, may allow access to this data (`http://www.niss.org/eia/niss-eia-microdata.html`). The dataset includes monthly observations for all refineries in the US on production, capacity, utilization, and inputs into production. The program is currently on hold.

available on EIA's website.[22] For every month in the years from 1995 to 2006, and for each of the 9 refining districts, I have the following data:

- Wholesale gasoline production, sales, and prices.
- Wholesale distillate production, sales, and prices.
- Crude oil first purchase price and inputs into refineries.
- The capacity utilization rate.

This provides 1,296 observations. I also have annual firm level data for the same years on the capacity to distill crude oil. The reported capacity, called the *atmospheric crude oil distillation capacity*, measures the number of barrels of crude oil that a refinery can process through the initial distillation process. This measure is calculated on a stream-day basis.[23]

There are 246 unique plants in the dataset, with 179 active in 1995 and 144 active in 2006. Overall, I observe a total of 1,959 plant-year observations. Table 1.2 summarizes the data by district and indicates the market definitions I use in my estimation. The number of plants and aggregate capacity are for January 2006.

Proceeding with the district level data on production and utilization combined with capacity at the firm level requires some discussion. Implicitly, I must make the

[22]See `http://tonto.eia.doe.gov/dnav/pet/pet_pnp_top.asp`. There are 9 refining districts, including the East Coast, the Midwest, the upper Midwest, the Central Plains, Louisiana, Texas, New Mexico, the Rockies, and the West Coast.

[23]Capacity reported in barrels per stream-day equals the maximum number of barrels of oil that a refinery can process on a given day under optimal operating conditions. Calendar-day capacities assume *usual* rather than optimal operating conditions, though these two numbers are frequently reported as identical.

Table 1.2: Industry Summary

Market	District	States	No. Plants	Ref. Cap. (Mbbl)
1	1	CT, DE, DC, FL, GA, ME, MD, MA, NH, NJ, NY, NC, PA, RI, SC, VT, VA, WV	14	659
1	2	IL, IN, KY, MI, OH, TN	14	913
1	3	MN ND, SD, WI	4	171
1	4	IA, KS, MO, NE, OK	8	306
1	5	TX	23	1,812
1	6	AL, AR, LA, MS	27	1,353
2	7	NM	3	42
2	8	CO, ID, MT, UT, WY	16	232
3	9	AK, AZ, CA, HI, NV, OR, WA	35	1,220
			144	6,709

strong assumption that all firms within a district are identical and respond the same way to shocks. When aggregating to the district, one firm that increases production may be cancelled out by another that breaks down. Thus, results from this approach will be meaningful only in terms of assessing the "average" behavior of a firm within a district. However, there is significant variation in district production levels as well as in the breakdown episodes described below. Also, aggregating to the district level when I estimate my model avoids having to account for the complicated linear programming problem faced by an individual refinery. These idiosyncratic differences should be smoothed out in the higher level data.

1.4 Model

Firms make annual investments to increase or decrease their available capacity. I assume these investments increase or decrease capacity immediately and that firms then choose their utilization rates each month. While empirically, some plants make major investments in capacity that take years to complete, the average investment is small and can be completely quickly.[24] Though plants require a certain minimum level of maintenance each year (usually carried out just before the summer driving season), running a plant at a high utilization rate in one month increases the probability of a plant breakdown or an extended maintenance episode in the next month. Thus, faced with relatively high product prices or low crude oil input prices (a high refining margin or *crack spread*), firms may want to run their plants at a high rate of utilization to maximize profits. However, this intensive use of capital may increase the possibility of a breakdown next month when prices may be even higher.

I model the competitive environment by assuming that plants are price-takers in the market for crude oil but are Cournot competitors with some (small) market power in the downstream refined products market. Since I do not observe plant level production choices, the model is best described as a representative-agent Cournot model. In each period, a firm optimally chooses its utilization rate in response to its estimate of the aggregate production of its competitors.

With the development of a network of pipelines across the US after World War II, markets tend to be large and feature many firms producing a homogeneous

[24] These small investments, known as *capacity creep*, include both additional infrastructure and improved through-put of existing capital.

product. Firms are differentiated not only by their capacity to turn crude oil into gasoline and other products, but also by their technical capabilities to utilize varying types of crude oil in their production. I focus on the capacity differentiation and average firm behavior to smooth over the technical production heterogeneity.

1.4.1 A Firm's Problem

Consider the problem of firm i in month m.[25] I will focus only on gasoline and distillate production by refineries, since these account for about three-quarters of the production of an average refinery. Denote production of gasoline and distillate as q^g_{im} and q^d_{im}, and the capacity of the refinery as $\overline{q}_{iy}$, where y indexes the current year. Given the investment behavior of firms, I assume that investments in capacity are made only once per year and the resulting capacity is fixed for the entire year. Let r_{iy} denote the investment of the firm, expressed as the proportional increase or decrease in capacity.

A firm's problem can be written as:

$$Max_{\{r_{iy}\}_{y=0}^{\infty}} \quad E\left[\sum_{y=0}^{\infty} \delta^y \Pi_{iy}(r_{iy}; x_{iy})\right], \tag{1.2}$$

$$\Pi_{iy} = Max_{\{u_{im}\}_{m=1}^{12}} \quad E\left[\sum_{m=1}^{12} \mu^{m-1} \pi_{im}(u_{im}; x_{im}, \overline{q}_{iy})\right]. \tag{1.3}$$

I assume capacity evolves according to:

$$\overline{q}_{iy} \quad = \quad \overline{q}_{i,y-1}(1 + r_{iy}), \tag{1.4}$$

[25] I assume that firms are individual plants and use the two terms interchangeably.

where r_{iy} is net of any depreciation of existing capital. The utilization rate can be expressed as:

$$u_{im} = \frac{q_{im}}{\overline{q}_{iy}}, \tag{1.5}$$

where $q_{im} = q^g_{im} + q^d_{im}$. While this is not a classic utilization rate, in that it does not assess the proportion of available inputs that are actively being used, technical constraints on the proportion of total capacity that can be used to produce gasoline and distillate makes this ratio approximately a scaled down version of the actual rate. $\pi_{im}(\cdot)$ is the per-period profit function, x_{im} and x_{iy} are vectors of state variables, and δ and μ are the discount rates, with $\delta = \mu^{12}$. Note that $\overline{q}_{iy}$ appears as a state variable in equation 1.3 and equals last year's capacity plus or minus the investment made at the beginning of the current year. Throughout a given year, state variables observable to the firm include the following:

P^c_{jm}	The price of crude oil
B_{im}	An indicator equal to 1 if the firm is in a breakdown episode
$Q_{-i,m}$	The estimated aggregate competing production by other firms in the market
$\overline{q}_{iy}$	A firm's capacity
$Time$	Month & year

I explicitly include a district j index on the crude oil price because, while I assume this price is exogenous, there are differences in the quality and price of oil in different

districts. The competing production state is needed to calculate the price of a firm's output. With the large number of firms in the industry, each firm has only a small impact on the prices of gasoline and distillate.[26] Firms form a statistical forecast of competing production as follows:

$$E[Q_{-i,m}] \quad = \quad Q_{-i,m-1}(1 + g_m), \tag{1.6}$$

where g_m is the historical growth rate of production in the market between months $m-1$ and m. The month of the year is included to capture the obvious and important seasonal effects. For example, a refinery operator may forgo preventative maintenance measures during the summer high-demand period to capitalize on the high prices and profit margins. The expectation operator is taken over the future profile of the state variables, some of which are deterministic (month and year), others of which evolve according to the firm's choices (capacity and breakdown), and still others are stochastic, for which firms base their expectations on historical values (the crude price and competing production).

Due to breakdowns, only a portion of $\overline{q}_{iy}$ will be available in a given month. I denote the *available* capacity as $\overline{q}^*_{iy}$. Because the numerator in equation 1.5 is the volume of downstream products and the denominator is the number of barrels of crude oil that a refinery can distill, the utilization rate may be greater than 1 in some cases. This occurs because chemicals called blending components are added in the distillation process (such as oxygenates like MTBE and ethanol).

[26] With plant-level production data, I could explicitly solve for the (asymmetric) Cournot equilibrium in each period. I plan to adopt this approach in future research.

Note that the firm's objective function can be written recursively. Denote $V(\cdot)$ to be the present discounted value of the stream of refiner's profits with optimal choices. Then, after dropping subscripts and discretizing the state space, the Bellman equation can be written:

$$V(x) \quad = \quad Max_r\Big\{\Pi(r;x) + \delta \sum_{x'} V(x')P(x'|x,r)\Big\}. \tag{1.7}$$

Here $P(\cdot)$ is the annual probability transition matrix and it reflects the transition between average annual values of the state variables. To solve for $\Pi(r;x)$, I apply backward induction from December back to January. For example, the expected value of a refiner's aggregate discounted profit from July onward is:

$$W_6 \quad = \quad Max_{u_6}\Big\{\pi_6(u_6;x_6,\overline{q}) + \mu \sum_{x_7} W_7(x_7)P^*(x_7|u_6,x_6,\overline{q})\Big\}. \tag{1.8}$$

Here, $P^*(\cdot)$ is conditional on u and $\overline{q}$ because plants that do not invest in new capacity and choose to operate more intensively increase their probability of breaking down.

1.4.2 Per-Period Profit

Prices are determined at the market level, which I index by k. Per-period profit is defined as gasoline and distillate revenue less production costs and investment

costs. Thus, in month m, profits of firm i are:

$$\begin{aligned}\pi_{im}(u_{im}; P^c_{jm}, , B_{im}, Q_{-i,m}, \overline{q}_{iy}, m, y) &= u_{im}\overline{q}^*_{iy}[(yield^g)P^g_{km}(Q^g_{km}; m, y) \\ &+ (1 - yield^g)P^d_{km}(Q^d_{km}; m, y)] \\ &- C_{im}(u_{im}; P^c_{jm}, \overline{q}^*_{iy}) \\ &- \frac{1}{12}C^r_{iy}(r_{iy}),\end{aligned} \tag{1.9}$$

where,

$$\overline{q}^*_{iy} = \begin{cases} \overline{q}_{iy} & \text{if } B_{im} = 0 \\ \phi\overline{q}_{iy} & \text{if } B_{im} = 1. \end{cases} \tag{1.10}$$

The term $yield^g$ represents the proportion of available capacity that can be distilled into gasoline. It is fixed over time and across firms. Functional forms for the demand and cost functions will be specified below. The last term in the profit function is the investment cost, which is spread equally across the 12 months of a year. Note that $\phi \in [0, 1)$ reflects the percentage reduction in capacity that a refinery experiences during a breakdown. While I allow this term to vary stochastically, the data suggest this value averages around 0.9 and can fall as low as 0.7. In other words, district level breakdowns occur that result in a 30% reduction in capacity relative to normal levels. It should be noted that a 25% capacity reduction in a given month could result from one week of complete breakdown and three weeks of optimal operation.

1.4.3 Demand

The prices of gasoline and distillate are determined at the "market" level. The three markets defined earlier are: the East Coast, Midwest and Gulf Coast; the Rocky Mountain region; and the West Coast. The first is by far the largest, with several large pipelines connecting the major production area near the Gulf of Mexico with the population centers on the East Coast and in the Midwest. I estimate the demand for wholesale gasoline (and similarly for distillate) according to:

$$log\ Q^g_{km}(P^g_{km}) \quad = \quad \alpha^g_0 + \alpha^g_1(Month) + \alpha^g_2(Year) + \alpha^g_3(log\ P^g_{km} * Year) + \epsilon^g_{km} \tag{1.11}$$

P^g and Q^g are the price and sales of wholesale gasoline. Here I specify a log-linear demand equation with month and year fixed effects to account for the strong seasonal variation and the growth in demand over time. I allow the price elasticity of demand to vary by year to account for the changes in the sensitivity of consumers to prices.

Note that the East Coast receives a significant amount of their refined product from abroad (mostly from Europe and the Caribbean). Imports increase in periods of high demand or tight supply, as the price must be high enough to justify the transportation costs. Thus the demand for refined products from US refineries may be affected by the availability of imports, though robustness checks reveal that the effect is small relative to the size of the East Coast's overall market (which includes the Midwest and Gulf Coast).

1.4.4 Probability of Breakdown

Consider the following specification for the likelihood of a plant breakdown or extended period of maintenance beyond the regular minimum level:

$$Pr(\text{breakdown in month } m) = F(\beta u_{i,m-1}) \quad = \quad \frac{exp(\beta_0 + \beta_1 u_{i,m-1})}{1 + exp(\beta_0 + \beta_1 u_{i,m-1})}, \tag{1.12}$$

which assumes the probability follows the logistic distribution. The same specification is used to model the likelihood that a plant recovers from a breakdown next period, conditional on being broken down this period. With more detailed firm-level data, an ordered probit may be the ideal specification, as it would account for both the magnitude and length of the breakdown episode. Modeling the breakdown dynamics based solely upon last month's utilization rate, and not, say, the average rate over the last six months, is primarily a computational simplification. The results using only last month's utilization rate are robust to other specifications.[27] See below for how I define a breakdown using district-level production data.

1.4.5 Production and Investment Costs

I assume the following production cost specification:

$$C_{im}(u_{im}; P^c_{jm}, \overline{q}^*_{iy}) \quad = \quad \gamma_0 * q_{im} + \gamma_1 * q^2_{im} + \gamma_2 * q_{im} * P^c_{jm}, \tag{1.13}$$

[27] Specifications involving the prior 3-month average rate or last month's deviation from historical rates yielded similar results. With firm-level data on production, one could also include the age of the refinery and perhaps the length of time since the last significant maintenance period.

where $q_{im} = u_{im}\overline{q}^*_{iy}$, the firm's actual production of gasoline and distillate in the current month.

I assume firms face increasing costs as they near their capacity constraint. To model this, I suppose firms have a quadratic production cost function and also include a term, γ_2, reflecting the major input of the refiner, crude oil. Refiners take this crude oil price as exogenous since the price is determined on the world market. As firms produce near their capacity, they may face increasing costs due to less time for maintenance, excess wear on their capital, and other effects that raise their marginal costs.

Investments in capacity are available immediately, and capacity is fixed within the year. This is a strong assumption since firms likely make investment decisions far in advance and spread the costs over a long time period. In future work, I will relax this assumption, allowing for a one-year "time-to-build." Investments come at a cost:

$$C^r_{iy}(r_{iy}) \quad = \quad \gamma_3(\overline{q}_{i,y-1}r_{iy}) + \gamma_4(\overline{q}_{i,y-1}r_{iy})^2. \tag{1.14}$$

The parameters, γ_3 and γ_4, reflect the cost of capacity expansion. They embody both the cost of physical expansion and any regulatory costs faced by the plant. Unfortunately, I will not be able to differentiate these two components with currently available data. Note that the investment cost parameters reflect the cost of a change in the number of barrels of a capacity that is created or destroyed. Large plants may benefit from economies of scale in capacity expansion as compared with

smaller plants, but since I am estimating my model for an average capacity firm, this consideration is not necessary.

1.5 Empirical Estimation Strategy

In general, I split the estimation into two stages. I first estimate the demand parameters, $(\alpha_0^g, \alpha_1^g, \alpha_2^g, \alpha_3^g, \alpha_0^d, \alpha_1^d, \alpha_2^d, \alpha_3^d)$, via GMM. This is a static relationship between the market price and quantity. I also estimate the logit parameters governing the probability of breakdown, (β_0, β_1), via maximum likelihood.

In the second stage, I take the demand and breakdown coefficients as given and solve the firms' dynamic utilization and investment choice problem using a *nested fixed-point GMM algorithm* to recover the cost parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ for each market. I allow for the cost parameters to vary each year to reflect changes in technology over time. I assume an annual discount rate of $\delta = 0.95$, implying a monthly rate of $\mu = 0.996$. When a firm enters a breakdown episode, I assume their capacity is reduced by a random amount, ϕ, which follows a beta distribution with mean 0.9.[28]

The firms' dynamic problem can be thought of as a finite-horizon monthly utilization choice problem nested inside an infinite-horizon annual investment choice problem. The annual investments in capacity can raise or lower the optimal utilization rate throughout the year, (e.g., a larger investment allows for the same level of

[28]Formally, $\phi \sim \mathcal{B}(9, 1)$.

output with a lower level of utilization). Recall that the problem can be written:

$$Max_{\{r_{iy}\}_{y=0}^{\infty}} \quad E\left[\sum_{y=0}^{\infty} \delta^y \Pi_{iy}(r_{iy}; x_{iy})\right], \tag{1.15}$$

$$\Pi_{iy} = Max_{\{u_{im}\}_{m=1}^{12}} \quad E\left[\sum_{m=1}^{12} \mu^{m-1} \pi_{im}(u_{im}; x_{im}, \overline{q}_{iy})\right]. \tag{1.16}$$

The aggregate discounted profits of the firm over the course of the year becomes the per-period (annual) payoff of the investment choice problem. Given the frequency with which refiners adjust their capacity and their utilization rate, this modeling strategy is not only realistic, but it is computationally appealing. Solving the finite horizon problem in equation 1.16 is simply a matter of backward induction.

The state variables available to the firm are the same in both sub-problems, aside from the month of the year, which is only relevant in the utilization choice problem. For the annual investment choice, the firm considers the average values of last year's crude oil price and market production, the proportion of time the refinery was broken down in the last 12 months, and the current level of capacity.

1.5.1 Demand

The demand parameters, the α's, are estimated in the first stage using 2-stage least squares with appropriate instruments. Given the endogeneity of P and Q, I need to find instruments, Z_{km}, that are correlated with the price, $Cov(P_{km}, Z_{km}) \neq 0$, and unrelated to error term, $Cov(\epsilon_{km}, Z_{km}) = 0$.[29] An obvious cost shifter in

[29]Essentially, I need cost shifters that move around the supply curve to trace out a demand curve.

the oil refining industry is the price of crude oil, which should be exogenous as it's determined in the world market. However, it is likely that the market for crude oil and the market for refined products are both subject to the same demand shocks, which invalidates the contemporaneous crude oil price as a good instrument.

Therefore, I instrument for the price of wholesale products with the lagged crude oil price, indicators of supply disruptions (such as those caused by hurricanes and pipeline outages), and the inventories of gasoline, distillate, and crude oil. These are industry-wide inventories, not just at the refinery. These should all be related to the price of a refiner's products though unrelated to the downstream demand. I can use the R^2 from the first stage to test for the correlation between my instruments and the endogenous price. Since I have instrumented for price in the first stage, in the second stage I regress the log of Q_{km} on the fitted log price, along with year and month fixed effects.

1.5.2 Breakdown Probability

The parameters of the breakdown logit, β_0 and β_1, are estimated by maximum likelihood. This is done separately for estimating the likelihood that a breakdown occurs and for the likelihood that a plant recovers once broken down. I define a "breakdown" in district j as a month when the observed utilization rate u_{jm} (published by EIA, reflecting gross inputs of crude oil divided by the capacity to

distill crude oil) drops below $\underline{u}_{jm}$, defined as:

$$\underline{u}_{jm} \quad = \quad min\left\{\frac{1}{9}\sum_{i=1}^{9} u_{im}, \frac{1}{4}\sum_{i=1}^{4} u_{j,m-12i}\right\}.$$

So the threshold is the smaller of the contemporaneous average across all districts and the average of the selected district's production in the *same* month for the last 4 years. So a breakdown is only triggered when 1) a district is producing relatively less than all other districts in the current month, and 2) the district is producing relatively less than it has historically in the same month. Figure 1.8 displays the breakdown dynamics for districts that experience a breakdown. The plots show that districts that run their plants more intensively in one month are more likely to break down the following month.

Once a breakdown episode is started, a district may stay below the threshold for a period of months. The data show that median episode length is 1 month, the mean is 2.3 months, and the maximum is 15 months.[30]

1.5.3 Production Cost Parameters

The cost parameters, $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$, are estimated by GMM in the second stage dynamic optimization. In order to solve for the production and investment cost parameters, I need to solve a dynamic optimization problem. To achieve this,

[30]The 15 month episode occurred in district 9 (the West Coast) from February 1999 - May 2000. It resulted from two California refinery fires at the Tosco Refinery in Avon on 02/23/99 and at the Chevron Refinery in Richmond on 03/25/99. The fall in gasoline production from these two fires was only 7% but due to California's strict environmental standards for gasoline, shipments from other (less regulated) districts were impossible so prices rose by about 25%. This implies a demand elasticity for retail gasoline of -0.28.

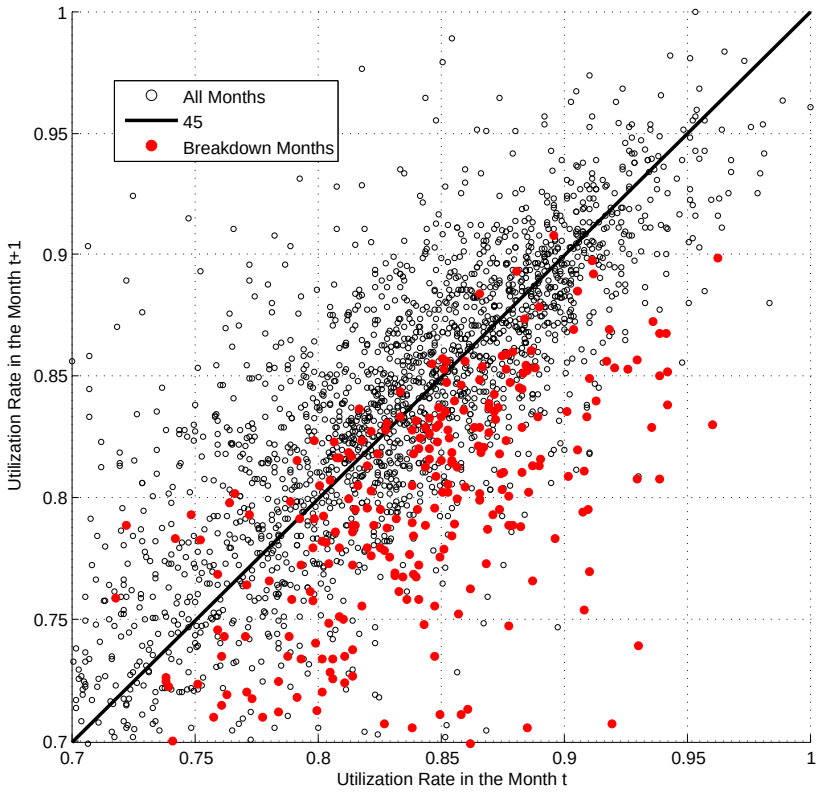


Figure 1.8: District Breakdowns

I first discretize the state space, which includes deterministic time states. The transition probability for the crude price is found using the empirical distribution of its historical series. The transition probabilities between breakdown states depend on the choice variable in the previous period according to the logit estimation done in the first stage. In a given year, the transition matrix for months reflects moving from one month to the next with certainty. Therefore, I can simplify the analysis by taking advantage of the cyclic nature of the month state. This dramatically reduces the computational time; see Rust (forthcoming). Further details of the estimation algorithm can be found in appendix C.

For a candidate parameter vector, I iterate on the policy function until convergence. I then interpolate the policy function on the actual states in my data and estimate the utilization rate for each district-month observation. Since the optimization is preformed at the firm level, I aggregate to the market level and form the following moments:

$$\begin{aligned} M_1 &= J^{-1} \sum_j (u_{mj} - \hat{u}_{mj}) \\ M_2 &= N_j^{-1} \sum_i (r_{ijy} - \hat{r}_{ijy}) \end{aligned}$$

where $\hat{u}_{mj}$ is the *average* utilization rate in district j and month m and $\hat{r}_{ijy}$ is the estimated investment rate by firm i located in district j in year y. I average the utilization rate moments over districts and the investment rate moments over firms and then stack them to form a moment vector: $M(\gamma) = (M_1, M_2)'$. I then

numerically solve the following problem:

$$Min_{\gamma} \quad \Big\{ M(\gamma)' \Psi^{-1} M(\gamma) \Big\}, \tag{1.17}$$

where Ψ is the variance-covariance matrix of the moment vector. With estimated parameters in hand, I estimate the standard errors of the cost estimates using Hansen's GMM estimator of the VC matrix. Given the matrix G of numerical derivatives, where (for parameter k and moment l)[31],

$$G_{lk} \quad = \quad \frac{M_l(\overline{\gamma}_k) - M_l(\underline{\gamma}_k)}{\gamma_k * 1\%}, \tag{1.18}$$

I can then compute:

$$VC(\gamma) = \frac{1}{N} (G' \Psi^{-1} G)^{-1}. \tag{1.19}$$

1.6 Results

1.6.1 Model Fit

I first assess the fit of the dynamic model by plotting actual and estimated values of key variables in figure 1.9. This is an in-sample analysis and shows that, on average, the estimated values match the data fairly well. Prices are estimated very precisely due to the flexibility gained by including monthly fixed effects. The estimated utilization rate is more variable than the actual rate though the month-

[31] For a 1% window, I perturb the parameter by 0.5% above and below the estimate.

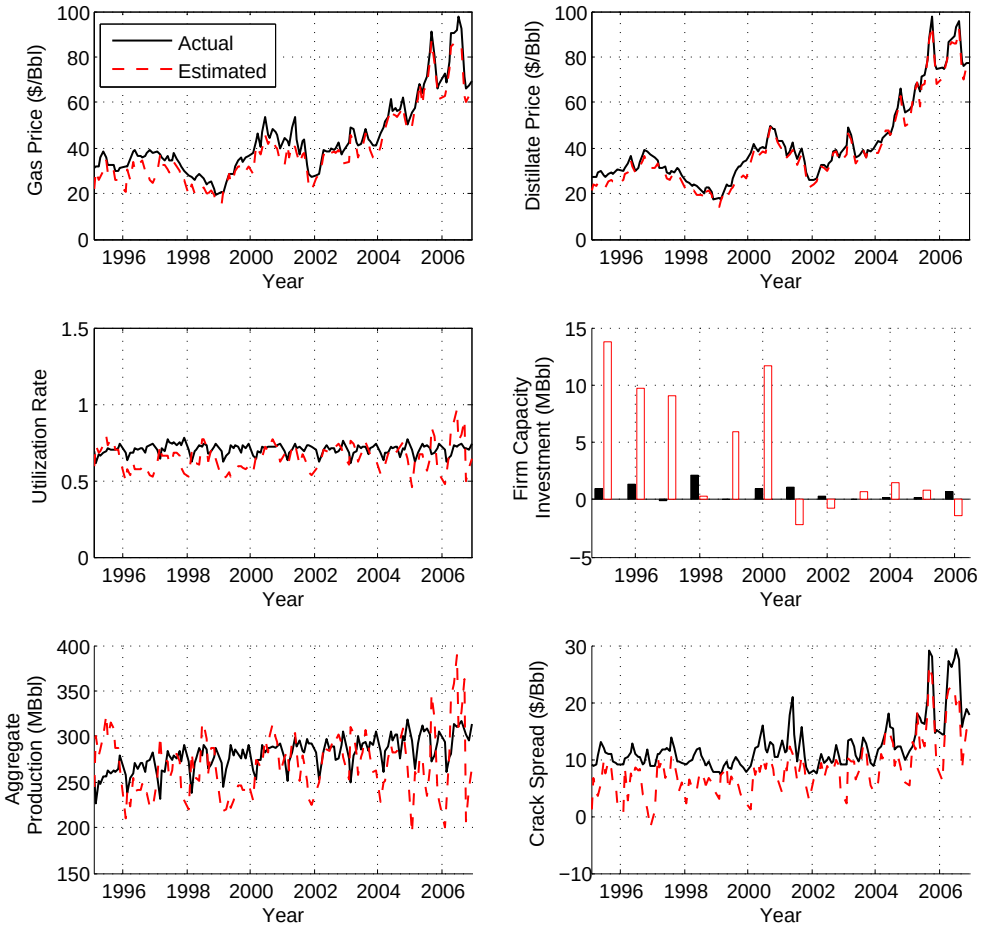


Figure 1.9: Model Fit (In Sample)

to-month fluctuations are approximated well. The model does not do as well at predicting the level of investment because firms tend to make lumpy investments every few years instead of updating their plant continuously. This means the median investment in any given year is zero and the reduced variation makes identification more difficult.

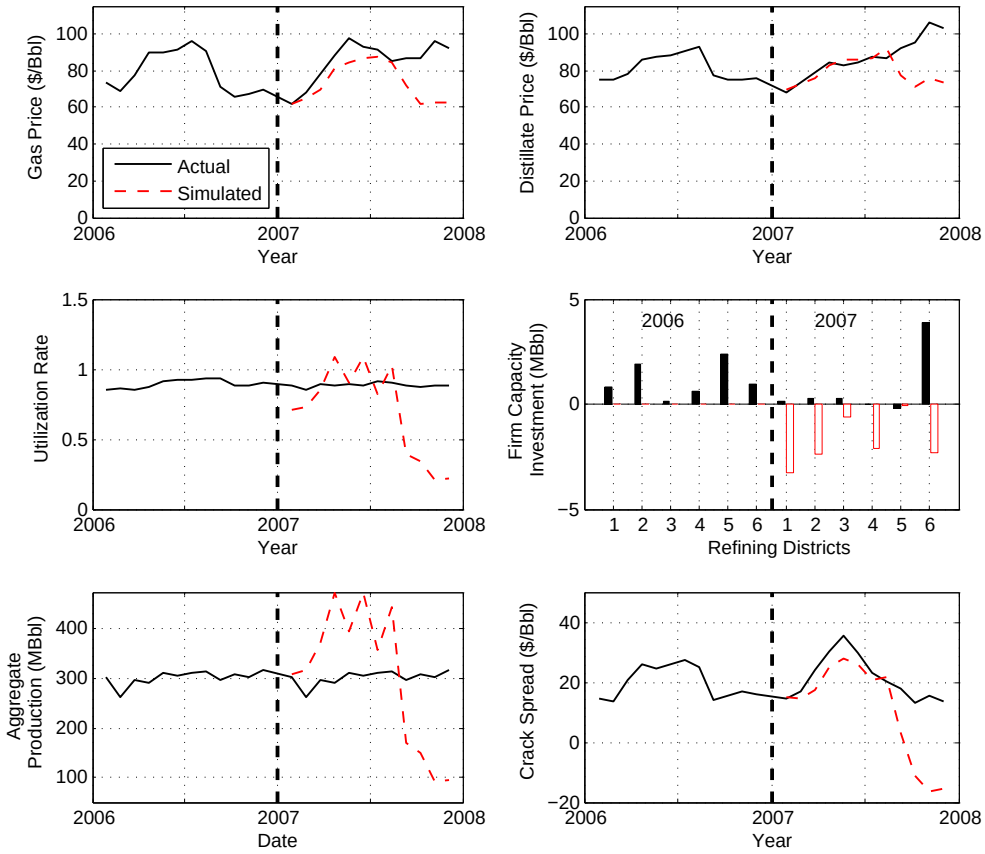


Figure 1.10: Model Fit (Out of Sample)

Finally, though the model tracks the movements in the crack spread very well, it tends to predict a value that is below the actual spread. This occurs because the estimated prices of gasoline and distillate are also biased down, because I do not account for inventories in my model. Since a small portion of refinery production is

stored, my estimates of downstream demand are biased up, which pushes down the estimated prices.

In figure 1.10, I do an out-of-sample test of the model, where I use the parameter estimates based on data through 2006 and simulate the investment and utilization policy of firms in 2007. The predicted prices of gasoline and distillate are close to the data for the beginning of 2007 but then begin to deviate. This pattern, also shown in the crack spread plot, is partially a result of unprecedented levels of the price of crude oil in 2007. The model predicts that refineries should optimally respond to these high input prices by cutting their utilization rate to drive up their product prices and maintain their profit margin.

1.6.2 First Stage Estimates: Demand and Breakdown

Tables 1.3 and 1.4 present the results of the first stage demand and breakdown estimations. Most of the demand coefficients are significant at the 1% level and have the expected signs. The monthly fixed effects estimates show the peak in gasoline demand during the summer months and distillate toward the fall. The elasticity estimates show a growing sensitivity to wholesale gasoline prices over the years. These estimates are higher than those reported for retail gasoline in other studies (see Knittel (2008)). However, unlike the branded retail product, wholesale gasoline is very homogeneous and downstream buyers can more easily substiute to a competing supplier. Also, the ability to store gasoline at terminals would imply the wholesale elasticity should be higher than the retail estimate. The R^2 from the

first stage regression of price on the instruments is 0.87.

The logit estimation of breakdown reveals an increasing probability of breakdown as a refiner runs the plant more intensively. Estimating the probability of breakdown next period conditional on being broken down this period reveals that refiners with more severe breakdowns are less likely to recover in the next period.

1.6.3 Second Stage Estimates: Costs

The cost coefficients are generally significant and reflect a production cost function that is increasing and convex. I display the cost functions at the average values of the estimates in figure 1.11 and report all estimates in appendix D, table A.3. The cost functions show that firms in market 2, the isolated Rocky Mountain region, are the most sensitive to production changes and have the highest overall production costs. Market 1 enjoys relatively easy access to crude supplies in the Gulf region and has the lowest production costs. The curvature of the production cost functions shows that refiners face increasing marginal costs as they approach the limitations of their capacity. I use a constant crude oil price of $50/bbl in my estimated production cost function.

The estimates of investment cost functions reflect an almost linear relationship, with the quadratic term often insignificant. While the figure shows the average investment costs over time, table A.3 displays the increase in expansion costs that refiners have faced in recent years. The Senate's (2002) estimated cost of building a new 2,700 barrel/day refinery was about $27 million. I estimate the cost of the

Table 1.3: Demand Estimates

	Gasoline		Distillate	
Parameter	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**
Constant	-0.27	0.44	5.20***	1.73
Year '95	2.51***	0.60	-0.99	2.43
Year '96	2.64***	0.64	0.93	2.70
Year '97	3.49***	0.62	0.90	2.55
Year '98	3.08***	0.56	-1.38	2.28
Year '99	3.44***	0.58	-1.82	2.35
Year '00	3.18***	0.67	-0.19	2.81
Year '01	3.19***	0.63	0.55	2.62
Year '02	3.59***	0.61	-0.09	2.47
Year '03	3.72***	0.65	1.58	2.75
Year '04	3.65***	0.65	0.52	2.84
Year '05	3.55***	0.65	1.33	2.96
Year '06	2.84***	0.61	1.24	2.86
Month 2	0.05***	0.01	0.04	0.04
Month 3	0.11***	0.01	0.13***	0.04
Month 4	0.17***	0.01	0.18***	0.04
Month 5	0.22***	0.01	0.14***	0.04
Month 6	0.25***	0.01	0.14***	0.04
Month 7	0.25***	0.01	0.11***	0.04
Month 8	0.28***	0.01	0.27***	0.04
Month 9	0.21***	0.01	0.33***	0.04
Month 10	0.19***	0.01	0.39***	0.05
Month 11	0.13***	0.01	0.26***	0.04
Month 12	0.11***	0.01	0.13***	0.04
Log(P)*Year '95	-0.81***	0.13	-1.79***	0.57
Log(P)*Year '96	-0.81***	0.14	-2.25***	0.64
Log(P)*Year '97	-1.07***	0.13	-2.27***	0.59
Log(P)*Year '98	-1.03***	0.12	-1.73***	0.52
Log(P)*Year '99	-1.08***	0.12	-1.50***	0.53
Log(P)*Year '00	-0.91***	0.14	-1.76***	0.63
Log(P)*Year '01	-0.93***	0.13	-2.02***	0.58
Log(P)*Year '02	-1.06***	0.12	-1.90***	0.54
Log(P)*Year '03	-1.04***	0.13	-2.25***	0.61
Log(P)*Year '04	-0.96***	0.12	-1.80***	0.59
Log(P)*Year '05	-0.88***	0.12	-1.82***	0.58
Log(P)*Year '06	-0.69***	0.10	-1.74***	0.53

****, **, * Significant at the 1%, 5%, and 10% level respectively. Dependent variables: log of gasoline and distillate sales. First stage regression of price on hurricane and pipeline disruptions, lagged crude oil price, and stocks of crude oil, gasoline and distillate.*

Table 1.4: Breakdown Probability Estimates

	Conditional on No Breakdown		Conditional on Breakdown	
Parameter	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**
Constant (β_0)	-2.40***	0.44	0.91**	0.45
Utilization_{t-1}(β_1)	0.74	0.62	-4.03***	0.67

*Maximum likelihood estimates. ***, **, * Significant at the 1%, 5%, and 10% level respectively. Dependent variable = breakdown indicator.*

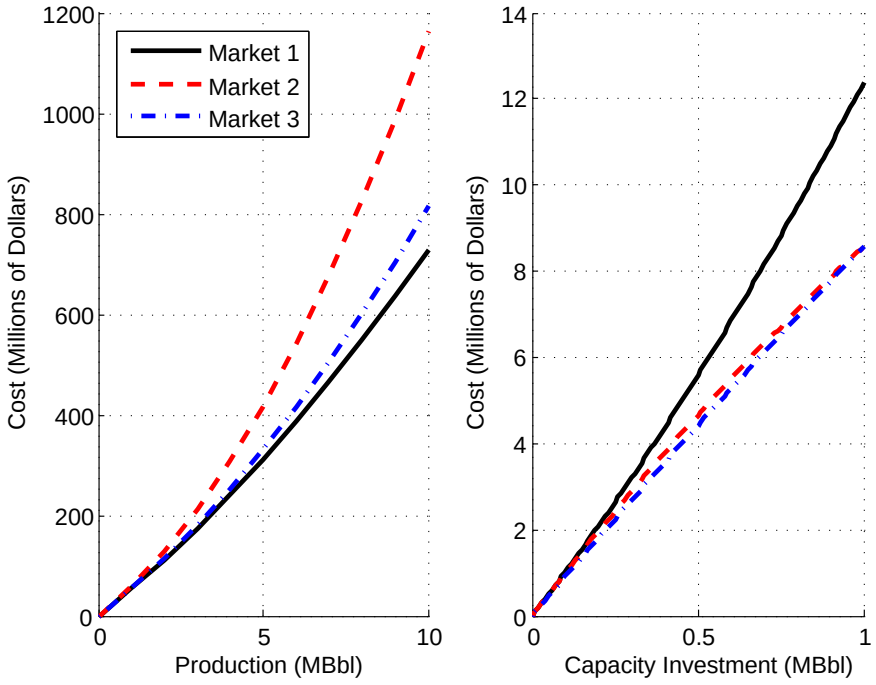


Figure 1.11: Estimated Production and Investment Cost Functions

same size *expansion* at around $10 million, further evidence that expanding existing sites is more cost-effective than building a new plant.

1.6.4 Policy Function

In figure 1.12, I plot the optimal policy function over the course of a year at the average values of the other state variables. The optimal utilization rate increases during the late winter and early spring but then falls off around April and May, before rising again to a peak in August. A likely explanation is that refiners, anticipating the high demand summer driving season in July and August, scale back operations in the late spring to prevent the possibility of a breakdown occurring during the peak. This pattern is replicated in most markets and years. Figure 1.13 displays the optimal policy function in 3-dimensional space, varying by

both the month of the year and the crude oil price. It shows that refiners cut back production when the oil price rises, a competitive response to a rising input price. The pattern across months is replicated at each crude oil price.

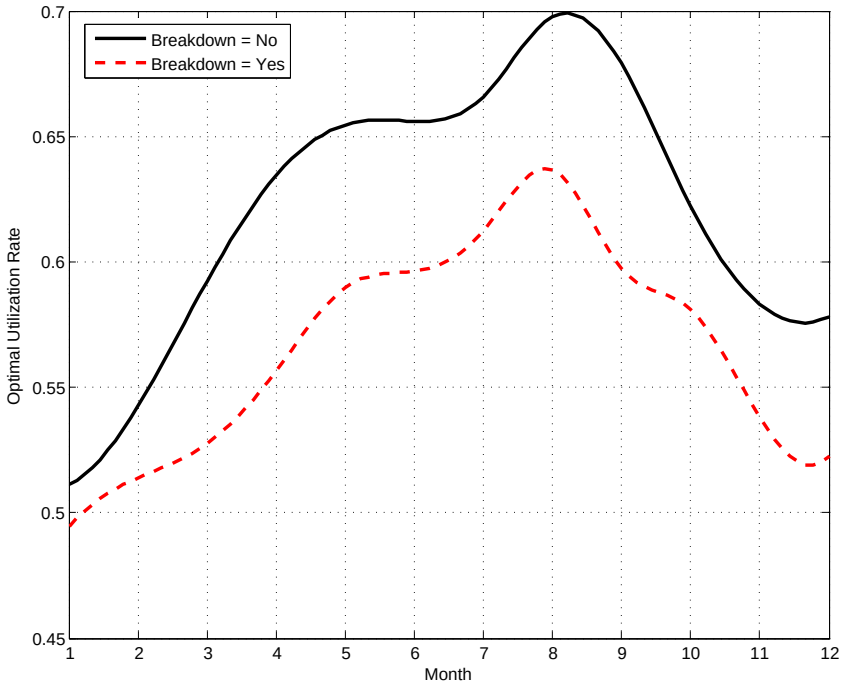


Figure 1.12: Optimal Utilization Rate Versus Month

1.7 Counterfactuals

With a fully estimated dynamic model of the US oil refining industry, I can now use the model to determine the effects of various shocks that may occur. There are many interesting questions that could be examined with my model given the importance of oil refining in US and global energy markets. I focus on three stylized facts that I believe to be particularly important in the following analysis: crude oil prices are rising to unprecedented levels; there is little to no excess capacity in the oil refining industry; and end-use consumers of refined products are becoming

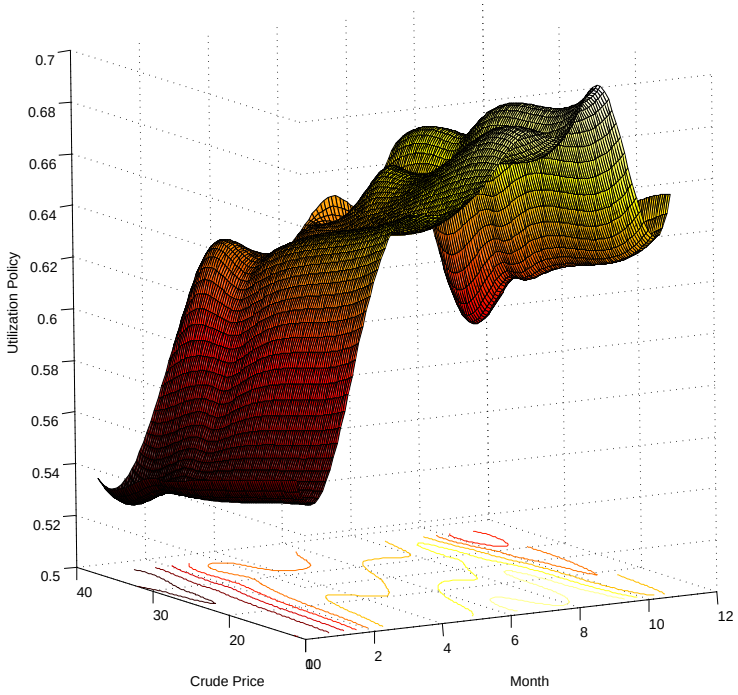


Figure 1.13: Optimal Utilization Rate Versus Month and Crude Price

increasingly sensitive to the prices they face (See Knittel et al. (2008)). Elasticities may be changing due to the availability of other fuels or because of changing perceptions of the environmental impact of oil usage (see figure 1.14). As a result, I will consider 2 experiments:

1. What are the effects of an increase in the crude oil price and how do the results change when the demand for refined products is more elastic?

2. What are the effects of a fall in available capacity and how do the results change when the demand for refined products is more elastic?

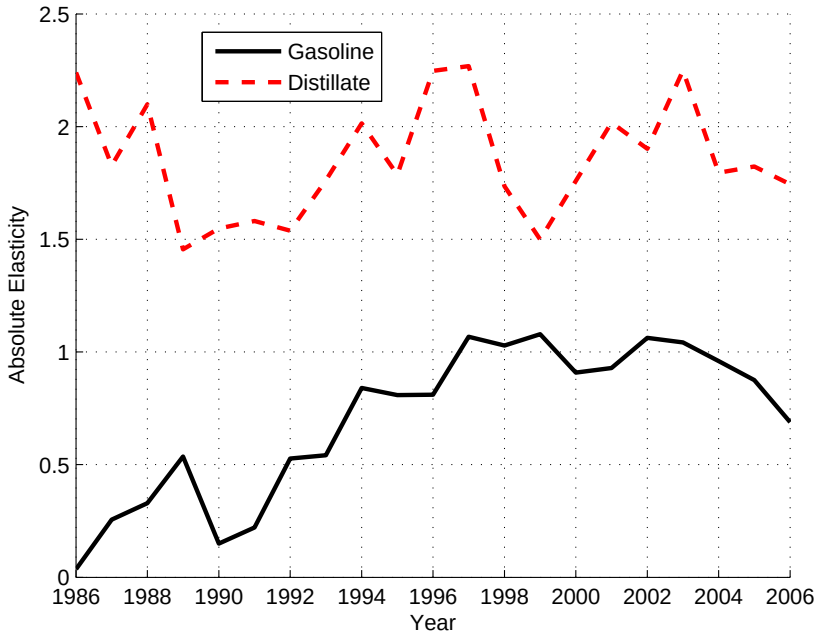


Figure 1.14: Price Elasticity of Demand

1.7.1 Methodology

Both counterfactuals are based on the coefficients and policy functions from 2006, the most recent year in my data. I shock the crude oil price in May to determine the effects throughout the peak demand summer months. The shock is permanent and I compute the average effects throughout the remainder of the year. I shock capacity in August to approximate the effects of a late summer hurricane hitting the Gulf of Mexico. I compute impacts assuming both the actual estimated elasticity in 2006 and an elasticity that is higher by 2.5% (in absolute terms) for both gasoline and distillate. Even this small increase in the sensitivity of consumers is enough to induce a dramatic response.

In my sample, the maximum observed real crude oil price is around $70/bbl. However, as shown in figure 1.15, crude oil prices have been driven to record levels

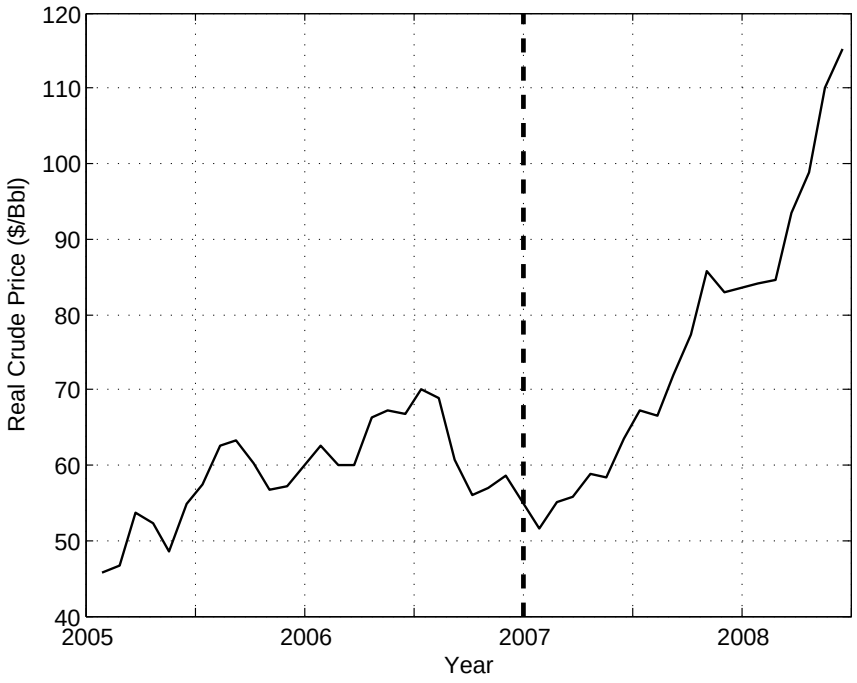


Figure 1.15: Crude Oil Price

more recently, exceeding \$115/bbl (in real 2006 dollars). Thus, I simulate the effects of a 20% increase in the price of crude oil to determine the impact on prices of gasoline and distillate and the resulting crack spread. Since the price elasticity of demand is one of the parameters estimated in the first stage and it influences the per-period payoff of the firm, I must solve my model at each new elasticity estimate. The optimal policy functions change as a result. Since the crude oil price is a state variable, I extrapolate my policy functions to the new crude prices.

About one-half of the US refining capacity is located on the Gulf of Mexico. Major hurricanes like Katrina and Rita in 2005, and more recently, Gustav and Ike in 2008, reduced US oil refining capacity by 25% to 35% and had a major impact on downstream prices and refiners' profit margins (see figure 1.16). Therefore, in my second counterfactual experiment, I simulate the effects of a 25% reduction

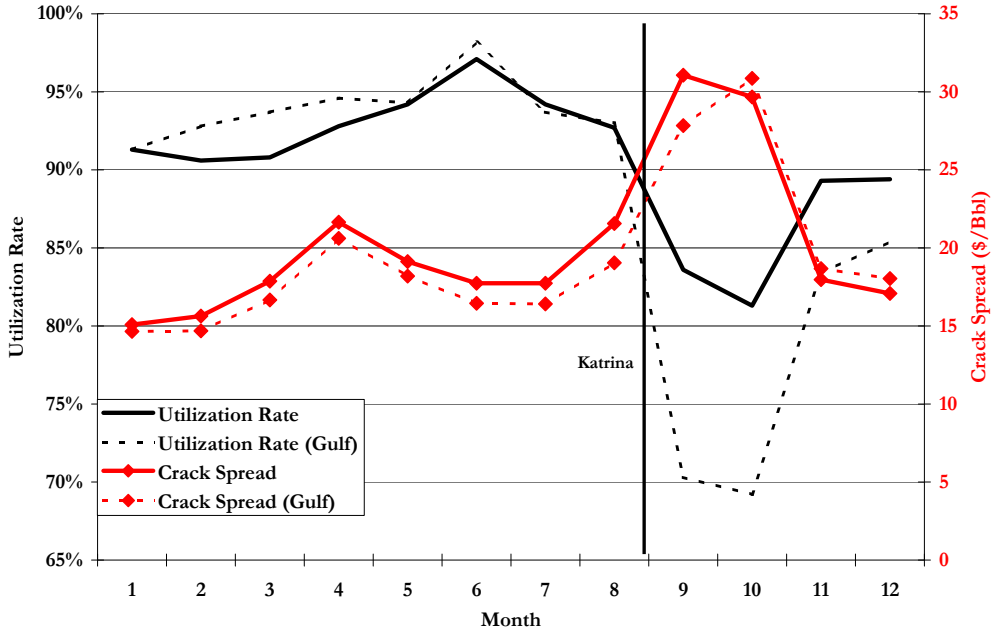


Figure 1.16: Loss in Capacity: Hurricane Katrina

in capacity on downstream prices, the crack spread, refiner profits, and consumer welfare.

1.7.2 Results of Experiments

The effect of a 20% increase in the price of crude oil (from 2006 prices) is shown in figure 1.17 and summarized in table 1.5. Note, the price and crack spread changes in the table are the average changes relative to the baseline prediction following the shock for the remainder of the year. The changes in surplus, profit and welfare are based on totals for the remainder of the year following the shock. The graphs in figure 1.17 show the future path of product prices, the utilization rate, and the crack spread through the remainder of the year.

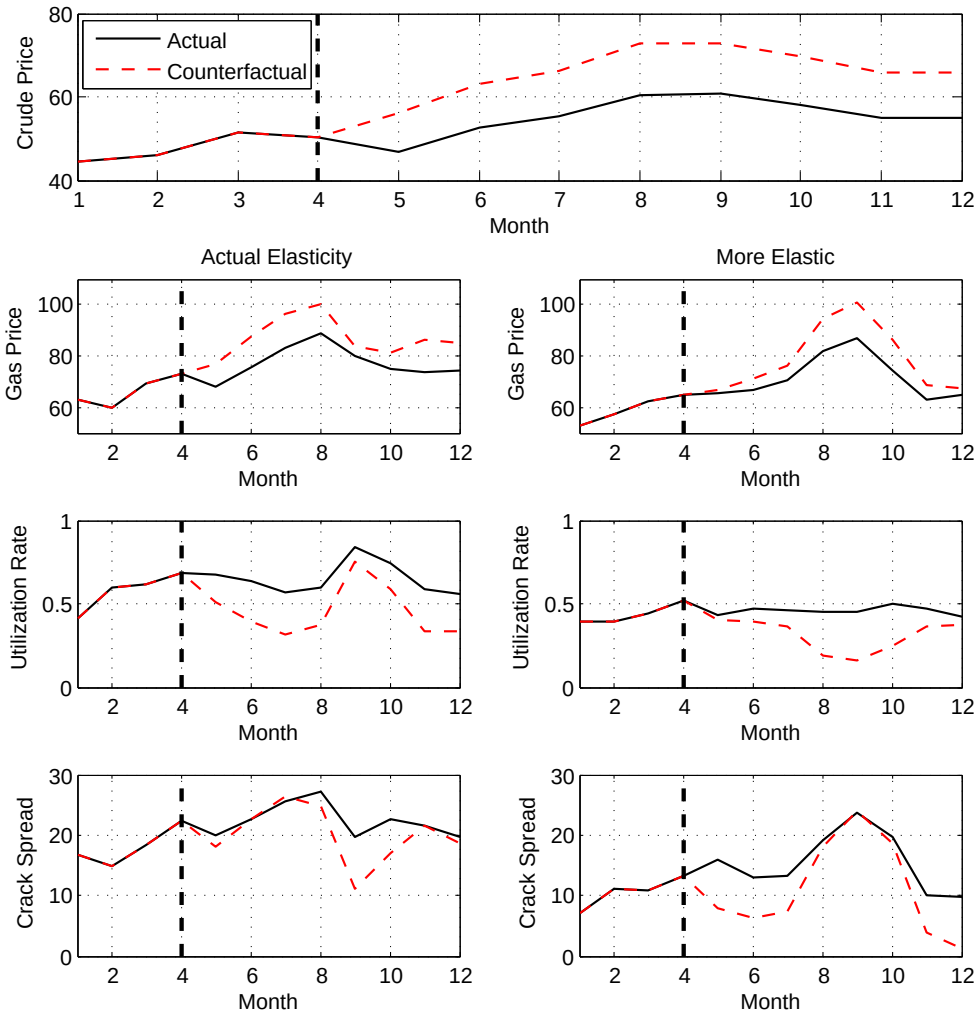


Figure 1.17: Crude Oil Counterfactual: Simulation

Table 1.5: The Effect of a 20% Increase in the Crude Oil Price

Percent Change	Actual Elasticity	More Elastic
Gasoline Price	12.7	10.2
Distillate Price	8.1	6.7
Crack Spread	-10.8	-30.1
Consumer Surplus	-58.3	-34.1
Refiner Profit	-37.1	-70.8
Total Welfare	-45.2	-49.7

The first column of graphs corresponds to the actual estimated elasticity (in 2006) and the second column of graphs assumes more sensitive demand estimates. The price of gasoline and distillate both rise following the crude oil price shock, though the price increases do not cover the entire cost increase as refiner profits fall after the shock. The amount of the increase that can be "passed through" to consumers appears to vary over the year. The crack spread graph reflects this, as it shows that although refiners are immediately hurt by the crude oil shock, they recover during the summer months by reducing their utilization rates before the spread falls again in September with weaker product demand.

Table 1.6: The Effect of a 25% Loss in Capacity

Percent Change	Actual Elasticity	More Elastic
Gasoline Price	15.9	3.0
Distillate Price	9.8	2.0
Crack Spread	47.9	11.9
Consumer Surplus	-69.0	-17.6
Refiner Profit	15.4	-4.8
Total Welfare	-11.1	-11.3

Comparing the two levels of demand sensitivity, we see that refiners are less able to pass on the crude price increase to more sensitive consumers, and thus their crack spread is dramatically reduced immediately following the shock. In addition to analyzing the effects on prices and profit margins, it is interesting to calculate the distribution of welfare between consumers and refiners. Total welfare declines by 45% in the months following the shock. According to table 1.5, overall welfare falls

for both the actual and more sensitive elasticity estimates, although more sensitive consumers end up with a larger share of the surplus following the shock.

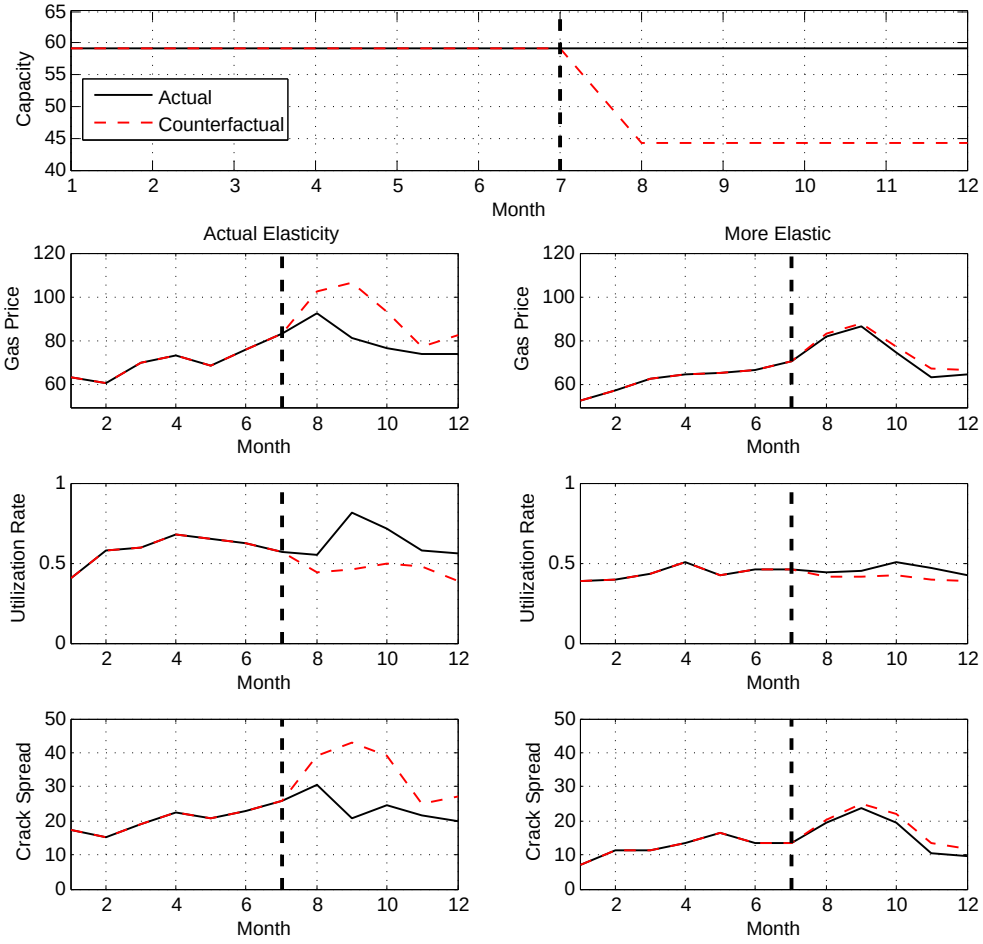


Figure 1.18: Capacity Counterfactual: Simulation

Figure 1.18 and table 1.6 display the results of my second counterfactual experiment, in which I reduce the size of the average refinery by 25%. Again, the table shows the average response to the shocks and figure 1.18 shows the longer-term effects for different levels of demand sensitivity. My counterfactual assumes that all refiners are hit equally hard by the shock, though in reality, some plants close completely while others operate even more intensively following events like Katrina.

The impact of the shock on the crack spread depends strongly on the demand elasticity. With the crude oil price the same in both cases and the percentage increases in the prices of gasoline and distillate about five times higher at the actual elasticity, the refiners facing more sensitive consumers benefit immediately following the shock, though the longer-term crack spread is higher for the less sensitive consumer group. Utilization rates change only slightly following the shock and the real cost is borne by consumers in the form of gasoline prices, which rise by almost 16%, reducing consumer surplus by 69%.

In terms of the distribution of welfare, the overall pie decreases by about the same amount in both cases, but at the actual elasticity, the increase in profits at operating refineries partially offsets the loss in consumer surplus. However, the more sensitive consumers retain a larger proportion of welfare following the shock. It's important to note that my measure of total welfare puts equal weight on consumer surplus and refiner profit and makes no consideration for the variability of prices faced by consumers. Given the economy's extraordinary reliance on gasoline, an extra dollar per gallon paid at the pump may hurt consumers more than it helps refiners.

1.8 Conclusion

In this paper, I have developed and estimated a new dynamic model of the US oil refining industry. Energy markets, and in particular, the production and distribution of gasoline, are a hot topic in both academic research and the popular

media. While the focus has tended to be on the upstream supply of crude oil (from both foreign and domestic sources) and the downstream retail stations, relatively little attention has been given to the role that oil refiners play in the industry. My analysis helps clarify and quantify the crucial role of the refiners in the transmission of crude oil and capacity shocks into downstream product prices, refiner profits, and consumer surplus.

The model matches the historical data and provides reasonably good out-of-sample predictions of key variables. I show that refiners are only partially able to pass through crude oil shocks to consumers and this ability varies across months of the year. As consumers have become more sensitive to changes in the price of gasoline, refiners face an even tougher competitive environment. Capacity disruptions, such as those caused by hurricanes, increase industry profits because the resulting price increase outweighs the loss in profits caused by reduced production. The effect on overall welfare is negative, though fairly small because the large loss in consumer surplus is partially offset by a rise in refiner profits.

My analysis not only models the behavior of refiners and the role they play in an important energy market, it also may have policy implications regarding optimal environmental regulations. In conversations with refiners, I found that current regulatory policies regarding both the building of new plants and the expansion of existing sites is the main hurdle that managers face when making their investment decisions. Regulatory policies have, at the very least, contributed to the current situation where capacity is tight and small shocks can have large effects. Realizing the importance of production flexibility in the refining industry means that new poli-

cies must balance responsible environmental concerns with incentives for capacity investment to meet the growing demand for refined products.

There are many extensions to this work that could provide further insights into the industry, though some require access to plant-level data which the EIA is considering making available. While this paper only addresses the production and investment decisions of active firms, including the possibility of exit may improve the model. Firms would likely follow a cut-off rule, exiting if the expected discounted stream of future profits fell below some critical level. Another potentially important determinant of firm behavior in this industry is a refiner's relationship with upstream crude oil producers. Currently, 60% of refiners are part of an integrated oil company, and although they benefit from a consistent supply of their major input, they are also constrained by having to exhaust their partner's stream of crude oil before seeking other, potentially more cost-effective sources. Independent refiners tend to invest in technologies that allow them to utilize different types of crude oil more flexibly, though may suffer relatively more when there is a supply disruption. Modeling the decisions of each type of refiner and the interaction between the two could help clarify the role of these vertical relationships. I leave these extensions for future work.

Chapter 2

Consumer Search for Online Drug Information

2.1 Introduction

There is a growing availability of medicinal drug information on the internet. A consumer seeking this complicated information faces the additional hurdle that the providers, e.g., drug companies, government regulators, and informational websites, all may have different incentives for providing accurate and unbiased information. While consumers formerly relied on their doctor as the primary source of information about the drugs they were taking, now they increasingly turn to the internet.[1]

Use of the internet worldwide doubled between 2004 and 2008.[2] When consumers go online, they are more likely to start with a search engine as the number of internet users accessing a search engine grew 69% between 2002 and 2008.[3] Thus, it is clear that search engines like Google and Yahoo are an important gateway to the internet. Also between 2002 and 2007, spending on Direct To Consumer Advertisting (DTCA) for prescription drugs by pharmaceutical companies doubled, with a small but growing portion of the online spending via banner ads and paid search

[1] "In 2007, 56% of American adults – more than 122 million people – sought information about a personal health concern from a source other than their doctor, up from 38%, or 72 million people, in 2001." (HSC August 2008). According to another survey, "approximately 40% of respondents with internet access reported using the internet to look for advice or information about health or health care in 2001." (JAMA 2003).

[2] `http://www.allaboutmarketresearch.com/internet.htm`. 0.757 billion in May 2004 compared to 1.463 billion in June 2008.

[3] `http://pewinternet.org/pdfs/PIP_Search_Aug08.pdf`. Pew Internet and the American Life Project (2008)

advertising.[4] The pharmaceutical company GlaxoSmithKline spent $2.5 billion dollars on advertising in 2007, of which $29 million (1.1%) was online spending.[5] A policy initiated by the Food and Drug Administration (FDA) in 1997 allowed detailed drug information to move to the internet with only essential side-effects and information provided in a television advertisement. The following year, DTCA on television more than tripled.[6]

My goal is to determine how consumers search for information and what characteristics of their query may determine how they navigate through the engine's results. This analysis focuses on the click behavior of consumers using AOL's internet search engine. I look at searches for brand name prescription drugs and those for consumer electronics as a comparison group. There are many reasons that consumers search and, like drug queries, a search for an electronics product may be motivated by a desire for product information which may lead to a purchase decision. Restricting to a specific group of products also allows me to define search sessions, discussed below, which are more difficult to determine in the entire universe of search queries.

Within drug queries, I analyze the effects of DTCA, drug age, and other drug characteristics (such as drug class) on consumer search. Given that consumers search, I also analyze how they do it: how in-depth (length of a search session, number of clicks, session time) and which types of links they click (extensions, ranks). I also analyze the different *drill-down* behavior between drug and electronics

[4]Source: TNS Media Intelligence.
[5]Source: `www.Adage.com`. Note this does not include paid search advertising.
[6]Television DTCA increased from $168 million in 1997 to $613 million in 1998.

searches. This is the frequent practice by users of submitting a query, processing the results, which may include clicking on one or more links, and then revising their initial search query.

I focus on drug-related search because typical consumers have limited information about the drugs that they are taking or are thinking about taking. The information also has many dimensions such as efficacy, side-effects, and interactions with other medications. Consumers face a wide variety of information sources both online and offline. My study complements the analysis in Day (2006), which investigates how consumers process and understand drug information via offline DTCA, though I only consider their initial search. As a result, consumers' understanding of the information they find is only relevant for this study in how it affects the way they search. For example, if consumers are frequently unsuccessful in finding the information they seek on dot-com sites, they may be more likely to click on other extensions in future search sessions.

The remainder of this paper is organized as follows. In section 2, I provide a description of the data which includes click-through data from AOL, drug information from the FDA, and advertising data from TNS Media Intelligence. Sections 3 and 4 include a descriptive and regression analysis on which types of search results are popular with consumers and how DTCA affects online search behavior, both in terms of the frequency and the intensity of search. Section 5 concludes and provides some directions for future work.

2.2 Data

AOL Click-Through Data

I focus on search and click-through data from AOL which spans a period from March to May, 2006. The data come from AOL Research, who posted the data on the web for research purposes on August 4, 2006. Due the privacy concerns, AOL later removed their own link to the data, but it is still available for download on many other websites.[7] The data has been used to study several topics including the determinants of search and how social networks could improve search engine performance.[8] To ensure privacy protection, I do not use any information specific to individual users and only report aggregate statistics in this paper. The data are a representative sample of over 650,000 AOL users and includes an anonymous user id, a date/time-stamp, a search query, and if the user clicked on a result, the domain portion of the click-through URL and its rank.[9] An overview of this data can be found in Chowdhury et. al. (2006).

In this analysis, I use the term *query* for a search event, which may or may not be followed by a click-through on a subsequent search result. If a user submits a query, clicks on a result, and then returns to the same search page (e.g., by clicking back on her internet browser) and clicks again, two observations are reported in the dataset with the *same* time stamp. If a user clicks on a result on page one of the search results (ranks 1-10), and then moves to page two, two observations are re-

[7] See `http://www.gregsadetsky.com/aol-data/`.

[8] See `http://www.cond.org/applications/paper3.pdf` and `http://www.stanford.edu/~koutrika/res/Publications/2008_wsdm.pdf`.

[9] If a user clicked on the link `www.fda.gov/drug/warnings.html`, only `www.fda.gov` is reported.

ported, but with *different* time stamps. Only organic results and not sponsored/paid results are included in the AOL database.

Drug Information

To create the database of queries, I use the FDA's Orange Book, which includes all drugs that have been approved by the FDA and attributes of each. These include the drug's age (years since FDA approval), drug class (16 broad classes), drug type (prescription, over-the-counter (OTC), or discontinued), and an indicator if drug is the Reference Listed Drug (RLD).[10] I select queries appearing in the AOL database that contain an FDA brand name somewhere in the query (i.e., it may appear among other terms). Of the 23,390 drug brand names appearing in the FDA Orange Book, 514 appear in the AOL database and account for 65,038 queries.

Advertising Data

I also gather data on DTCA for each drug in the sample. I have monthly data from 1994 through 2008 from TNS Media Intelligence. In 2008, the data include 327 drugs and the advertising expenditure is broken down by media type. Figures 2.1 and 2.2 display the growth of DTCA over time and the distribution of 2008 spending across media types. The growth of total DTCA is clearly evident and although TV and magazine advertising accounts for over 95% of total expenditures, spending on the internet is a new and growing outlet. TNS only reports internet ad spending on display or banner ads which appear, for example, across the top of many websites and some search engines. It does not include spending on sponsored/paid search

[10] A drug is an RLD if it is used as the chemical standard when generic versions of the drug are developed. New drugs have to be "bio-equivalent" to the RLD to gain approval by the FDA.

results which is reported to be twice the size of display ad spending.[11]

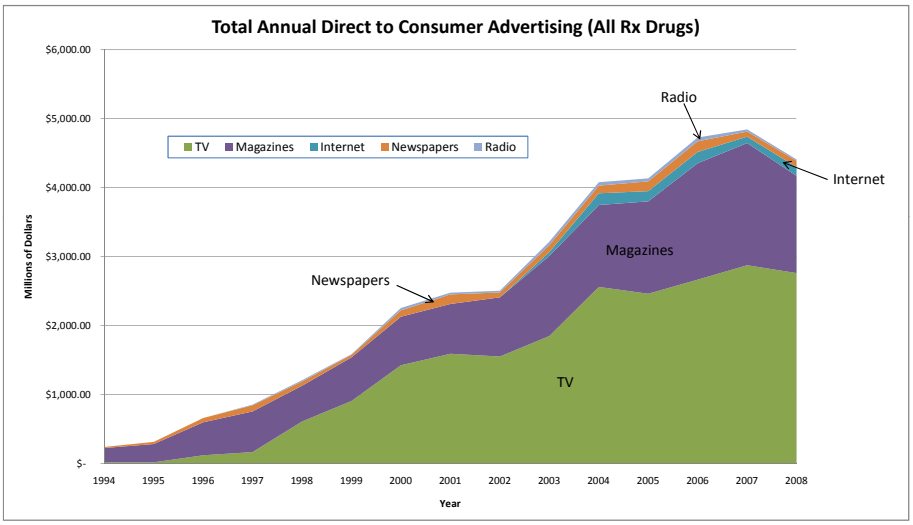


Figure 2.1: Total DTCA Spending on all Prescription Drugs

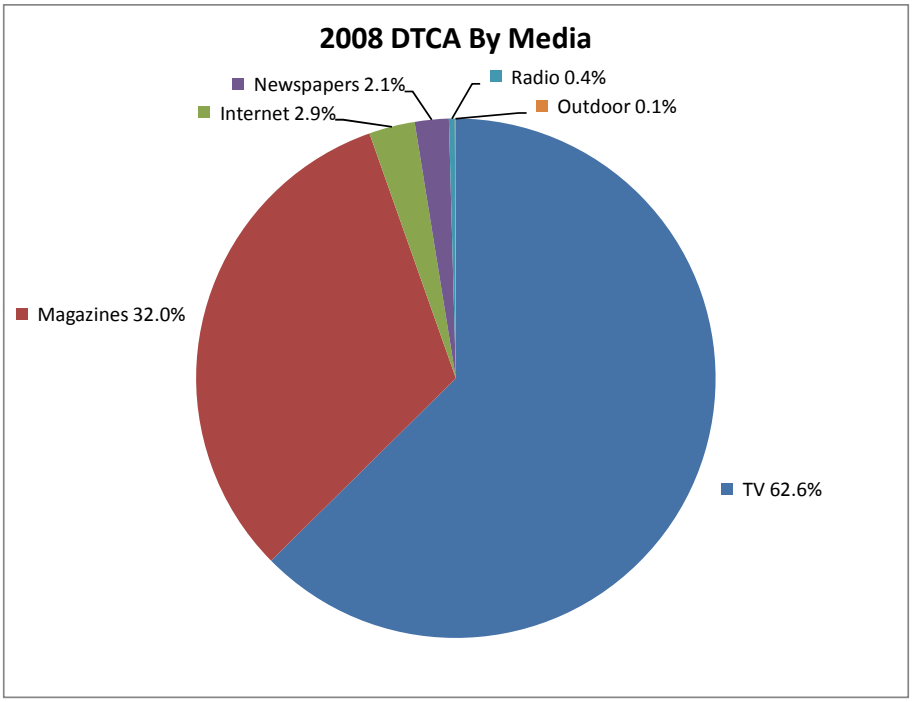


Figure 2.2: DTCA Breakdown by Media Type

Electronics

For electronics queries, I combined lists from consumer reports on popular electronics

[11] "Gap Widens in Online Advertising," *The Wall Street Journal*, September 4, 2008.

product and brand names with a list of manufacturers from tigerdirect.com, a major seller of consumer electronics. This resulted in 804 potential consumer electronics queries, of which 126 appear in the AOL database and account for 509,833 queries.

Generating Search Sessions

One challenge with analyzing search behavior on the internet is to group a sequence of potentially changing queries and click-throughs together to form a *search session*. Grouping identical queries together is frequently insufficient because users often revise their queries throughout a session. Therefore, I consider the following three approaches for defining a search session:

1. A sequence of queries with or without click-through such that the query is identical and the time between queries is less than one hour. The query needs to contain one of the drug brand names or electronics product words, but it may also include other words. However, the overall query may *not change* within a session which means the list of search results that the user sees is not changing. I use this definition for determining the popularity and transitions between website extensions and ranks.

2. A sequence of queries with or without click-through such that two adjacent queries are in the same session if any of the words appearing in the first query also appear in the second query. The time between queries is less than one hour.[12] I use this definition for the "All Queries" column of figure 2.1.

[12]There is a potential weakness in this definition. The three queries, "flights to Europe", "discount flights to London", "hotels in London" would all be classified in the same session though it is likely that the intent of the search changed in third query.

3. A sequence of "query-topics" with or without click-through where a keyword (such as a drug brand name) appears in all queries though other words may appear *and change* throughout the session. Again the time between adjacent queries must be less than one hour. This definition captures the drill-down behavior that users often exhibit when performing a search. I use this session definition for all other tables and regressions in the analysis.

2.3 Descriptive Analysis

Table 2.1 displays basic descriptive statistics of the AOL database including all queries and breakdowns for electronics and drug-related queries. Note that for the first column, I define sessions using method two, while for the drug and electronics sessions, I use method three. The method used for the all queries column is the most liberal in grouping adjacent queries together in a session, which should increase the number of multiple-query sessions. However, there are also many single-query sessions in the overall sample, which actually results in relatively fewer multiple-query sessions compared with drugs and electronics.

Compared with electronics sessions, drug sessions are more likely to feature multiple clicks, are shorter in time, though are longer in the number of clicks per session. As a result, the turn-over or average time between clicks is shorter for drug sessions than for electronics. This may be the result of a user in a drug-related session seeking a specific piece of information while an electronics session may involve a user attempting to get general information about a product. Electronics queries

Basic Statistics: AOL User Data

	All Queries	Electronics Queries	Drug Queries
Observations (Queries)	35,383,114	509,833	65,038
Click-Throughs	19,133,334	281,557	44,885
Num. Session	16,548,366	245,988	28,679
Users	651,559	110,261	17,459
Unique Query-Topics		126	514
Mean Users Per Query-Topic		875.09	33.97
Mean Queries per Session	2.14	2.07	2.27
Mean Session Length (Minutes)	2.71	3.55	3.00
Mean Time Between Queries	2.38	3.32	2.36
Multiple Query Sessions			
Num. Session	6,232,843	95,298	12,442
Proportion Mult Query Session	38%	39%	43%
Mean Queries per Session	4.02	3.77	3.92
Mean Session Length (Minutes)	7.19	9.16	6.90
Mean Time Between Queries	2.38	3.31	2.36

Table 2.1: Basic Statistics

are also dominated by several very popular terms as on average there are 875 users searching each query topic compared with only 34 for drug-related searches.

Table 2.2 is a breakdown of search activity and advertising by drug class.[13] The two largest classes account for over 42% of the search sessions but only 26% of the advertising spending. The lack of strong correlation between advertising and search is surprising if television ads (the largest component of DTCA) direct consumers to seek more information on the web. I will further investigate this relationship in the regression section below.

Two further slices of the data appear in tables 2.3 and 2.4 which show search and advertising activity by age and drug type respectively. Interestingly, though there is fairly high spending for younger drugs (1-3 years old), there is also relatively large DTCA on older drugs with the most on drugs that are 8 years old. Search

[13]Tables displaying the 20 most actively searched and advertised drugs in the sample can be found in tables B.1 and B.2 in the appendix.

Search Activity by Drug Classes

Drug Class	Class Num.	Num. Of Drugs	Num. Of Sessions	Num. Of Queries	Mean Queries Per Session	Ad Spending (Millions)
central nervous system agents	1	90	7,065	17,040	2.41	$666.70
psychotherapeutic agents	2	33	5,080	11,670	2.30	$220.17
metabolic agents	7	37	2,752	5,869	2.13	$474.76
anti-infectives	5	74	2,560	5,437	2.12	$92.69
cardiovascular agents	8	50	2,192	4,378	2.00	$32.24
miscellaneous agents	3	21	1,825	5,189	2.84	$353.92
hormones	6	48	1,709	3,996	2.34	$344.22
antineoplastics	9	36	1,299	3,113	2.40	$61.10
respiratory agents	10	21	1,283	2,514	1.96	$238.51
gastrointestinal agents	11	23	1,213	2,083	1.72	$417.57
topical agents	4	49	1,014	2,215	2.18	$452.17
coagulation modifiers	12	7	460	1,040	2.26	$110.21
not applicable	16	18	95	280	2.95	$0.10
nutritional products	14	5	91	151	1.66	$0.45
immunological agents	13	2	41	65	1.59	$0.17
biologicals	15	-	-	-		$0.00
Total		**514**	**28,679**	**65,040**	**2.27**	**$3,464.98**

Ad spending is total expenditure on all forms of DTCA in 2005.

Table 2.2: Search Activity by Drug Class

activity is fairly evenly spread among the different aged drugs though most activity is, again, on the 8 year old subset. The breakdown by drug type reveals far more activity on prescription drugs and even those classified as discontinued as compared with OTC drugs.[14] Non-innovator drugs, or drugs that are not designated as the RLD, receive a large share of both search activity and slightly more advertising spending.

Next, I analyze the search activity in the sample by looking at the popularity and transitions between various website extensions and ranks. Figures 2.3 and 2.4 display the percentage of clicks on each extension class of website and the percentage on clicks (within the first 10 clicks of a session) on each search result rank. I see that users in drug sessions click on relatively fewer dot-com results compared with electronics related searches. As expected, there is more attention paid to dot-gov and dot-org/net/info sites and this continues to grow in longer sessions (not shown).

[14]The advertising data only includes spending on prescription drugs (hence the $0 for OTC advertising spending), though in some cases I see spending on an OTC drug that has the same trade name as a prescription drug.

Search Activity by Drug Age

Age	Num. Of Drugs	Num. Of Sessions	Num. Of Queries	Ad Spending (Millions)
<1	35	609	1,368	$1.02
1	26	1,636	3,434	$442.72
2	30	1,171	2,668	$393.88
3	31	2,062	4,437	$287.30
4	25	1,064	2,481	$110.56
5	36	1,438	2,954	$263.13
6	25	1,555	3,503	$35.76
7	30	1,693	4,364	$8.73
8	31	2,771	6,706	$589.13
9	38	2,352	4,724	$317.41
10	23	1,285	3,262	$249.17
11	18	708	1,588	$115.27
12	17	1,110	2,344	$8.70
13	17	1,349	2,938	$323.53
14	16	1,424	2,976	$95.25
15	9	271	484	$5.30
16	6	280	561	$0.77
17	8	131	310	$0.22
18	8	570	1,273	$0.52
19	8	164	398	$0.32
20	9	566	1,389	$108.24
21	8	164	314	$0.61
22	1	384	1,127	$0.00
23	3	123	258	$0.00
24	3	316	867	$0.20
>24	53	3,483	8,312	$107.23
Total	**514**	**28,679**	**65,040**	**$3,464.98**

Ad spending is total expenditure on all forms of DTCA in 2005. Age is equal to years since FDA approval to March 2006.

Table 2.3: Search Activity by Drug Age

The rank popularity figure shows that attention by rank in drug sessions is less skewed toward the number one ranked sites. Users are more likely to click further down in the search results. The spike in the rank one popularity for electronics sessions is mostly driven by navigational searches (e.g., a search for apple.com and

Search Activity by Drug Type

Type	Num. Of Drugs	Num. Of Sessions	Num. Of Queries	Ad Spending (Millions)
Prescription	411	23,467	52,796	$3,163
Over-the-Counter	10	584	1,289	N/A
Discontinued	93	4,628	10,955	$302
Non-Innovator/RLD	267	18,023	41,214	$1,915
Innovator/RLD	247	10,656	23,826	$1,550

Ad spending is total expenditure on all forms of DTCA in 2005. An innovator is the original developer (pioneer) of the drug.

Table 2.4: Search Activity by Drug Type

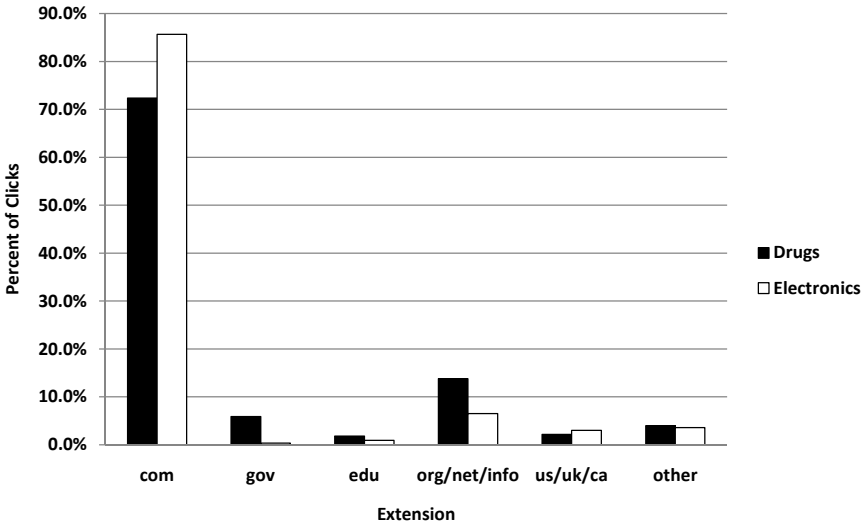


Figure 2.3: Extension Popularity in the First 10 Clicks

immediate click on the first search result, www.apple.com).

I also investigated how users make the transition between extensions and ranks in multiple click sessions. For tables 2.5 and 2.6, I generate sessions using method one to guarantee that the set of search results seen by the user on each click is identical. The table shows that transitions within extensions are less likely in drug search sessions compared with electronics sessions, though they both feature approximately the same exit rates. Transitions from other extensions to dot-gov and

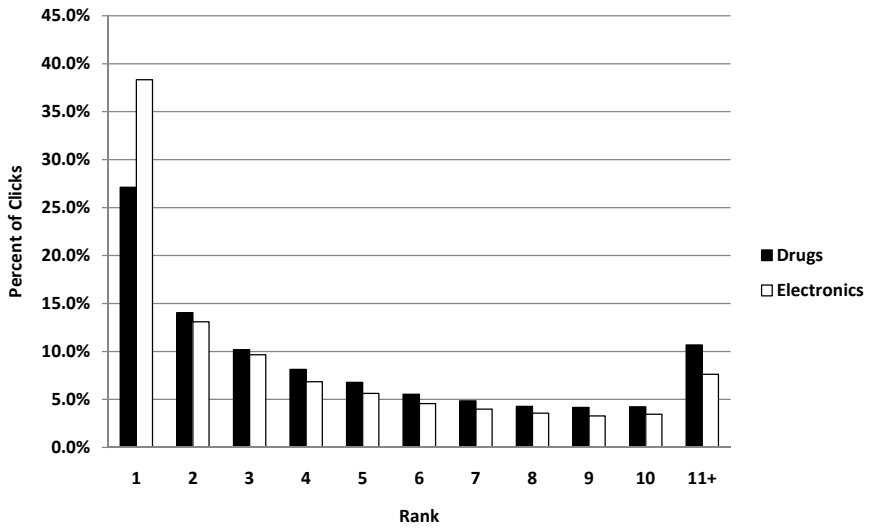


Figure 2.4: Rank Popularity in the First 10 Clicks

Drug Queries

EXTENSION (t)	EXTENSION (t+1) com	gov	edu	org/net/ info	us/uk/ ca	other	exit	
com	52.1%	2.9%	1.2%	9.6%	1.6%	3.0%	29.6%	100%
gov	41.2%	14.2%	1.2%	11.5%	1.5%	1.8%	28.6%	100%
edu	36.6%	3.1%	13.9%	12.3%	2.3%	4.4%	27.3%	100%
org/net/info	40.6%	3.6%	2.2%	20.7%	2.0%	2.9%	28.1%	100%
us/uk/ca	37.1%	3.3%	1.8%	11.7%	11.1%	4.8%	30.3%	100%
other	40.9%	1.6%	1.8%	9.0%	2.5%	17.1%	27.1%	100%

Electronics Queries

EXTENSION (t)	EXTENSION (t+1) com	gov	edu	org/net/ info	us/uk/ ca	other	exit	
com	61.0%	0.1%	0.5%	3.9%	2.0%	1.9%	30.6%	100%
gov	31.9%	20.8%	3.1%	10.2%	2.9%	1.5%	29.5%	100%
edu	36.5%	1.1%	22.8%	6.3%	3.6%	3.5%	26.3%	100%
org/net/info	42.8%	0.4%	1.1%	21.6%	2.4%	3.2%	28.5%	100%
us/uk/ca	49.0%	0.2%	1.2%	5.2%	14.8%	3.0%	26.7%	100%
other	39.7%	0.1%	0.7%	5.6%	2.6%	20.7%	30.6%	100%

These tables include the probability of transitioning from one extension to another during a search session. All search sessions are included as long as the user clicked on at least one link. A session is defined as a sequence of clicks following the *identical* query where the time between clicks is less than 60 minutes.

Table 2.5: Transitions between extensions

dot-org/net/info sites are more likely in drug searches.

The table on rank transitions reveals that users in electronics session are more likely to find what they are looking for on the rank one result as they are more

Drugs

RANK (t)	RANK (t+1) 1	2	3	4	5+	exit	
1	8.5%	26.6%	14.8%	9.4%	23.5%	17.1%	100%
2	11.2%	5.0%	19.6%	11.5%	26.4%	26.3%	100%
3	8.7%	6.1%	4.2%	18.2%	34.4%	28.4%	100%
4	6.8%	4.6%	4.5%	4.6%	49.0%	30.6%	100%
5+	3.5%	2.2%	1.7%	1.9%	56.3%	34.4%	100%

Electronics

RANK (t)	RANK (t+1) 1	2	3	4	5+	exit	
1	25.0%	17.4%	9.9%	5.9%	14.9%	27.0%	100%
2	12.5%	11.8%	17.6%	9.1%	21.0%	28.1%	100%
3	8.9%	6.3%	9.6%	16.6%	30.8%	27.7%	100%
4	7.0%	4.3%	4.9%	8.0%	46.0%	29.7%	100%
5+	4.3%	2.2%	2.1%	2.0%	55.9%	33.5%	100%

These tables include the probability of transitioning from one rank to another during a search session. All search sessions are included as long as the user clicked on at least one link. A session is defined as a sequence of clicks following the *identical* query where the time between clicks is less than 60 minutes.

Table 2.6: Transitions between ranks

likely to revisit it immediately (rank 1 to rank 1). They are also more likely to exit following a click on rank one. Given that sessions here are defined as unique queries, it may also be that electronics users are more likely to reformulate/refine their query after clicking on the rank one result which is classified as an exit. This turns out to be the case, as shown in the analysis of the potential for query reformulation, or "drill-down," in figure 2.5.

Several key features can be seen in the figure depicting drill-down behavior. Drug sessions are more likely than electronics session to involve a query followed by a click versus a query without a resulting click. Following a query with or without a click, users in drug sessions are more likely to issue the same query, less likely to

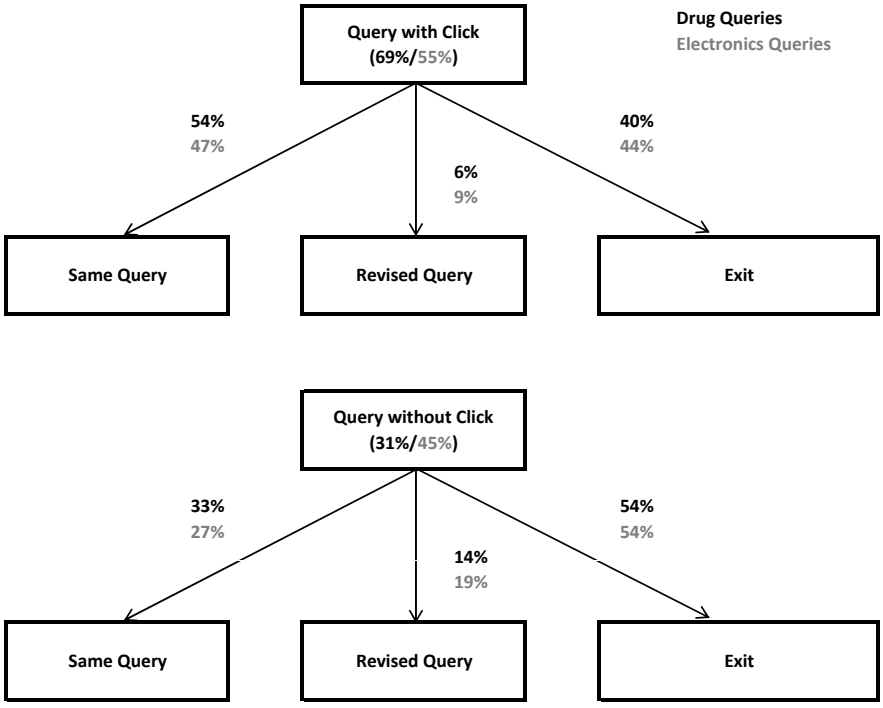


Figure 2.5: Drill Down Behavior

revise their query, and approximately equally like to exit as electronics users. If a user submits a query and clicks on a result, they are more likely to maintain the same query on the next click, less likely to revise and less likely to exit the search. It appears that query revisions are an important part of search behavior, but relatively more popular in electronics sessions compared with drug-related sessions.

Next, I turn to a more detailed analysis of consumers' search behavior where I investigate the determinants of both the frequency and intensity of drug related search. Without data on product attributes or advertising for consumer electronics products, the next section will focus only on drug-related search.

2.4 Regression Analysis

In this section, I report regression results explaining the determinants of consumers' search patterns. I look at drug-level regressions to determine how drug attributes affect search. Then in the session-level probit regressions, I determine how the intensity of search is affected by drug attributes and DTCA. A description of all variables included in these regressions is shown in table B.3 in the appendix.

2.4.1 Frequency Regressions

In the following set of results, I assess how DTCA and drug characteristics affect the frequency of search using drug-level data. I include the effects of both overall DTCA and also each individual media category that is available. This breakdown by media is important because DTCA via different media channels may have different effects on consumer search patterns. Television advertising, especially since the new FDA regulations in 1997 lessening the requirements on what needs to be conveyed during the ad, tends to only highlight the main benefits and potential side effects of a drug. Magazine ads usually include two pages: one with the highlights of the drug in full color and dramatic fonts, and the other with the details in fine print. The internet ads captured in the data are so-called *banner ads* and would likely have a similar effect to television DTCA with only the highlights presented. The same could said for radio and outdoor ads while DTCA in newspapers is likely presenting similar information to magazine ads. Therefore television, internet, radio and outdoor ads, given their lack of detailed information, may have a stronger

positive effect on search compared to magazines and newspapers.

Since I only observe DTCA on prescription drugs, I restrict the analysis by excluding OTC drugs. I investigate the effects of a drug's age on search as well as if the drug is designated as the RLD. Table 2.7 displays the results of a regression where the dependent variable is "total sessions," or the total number of search sessions I observe in the dataset over the three-month period for a given drug. Sessions are defined using method three, so they allow for keywords to change throughout the session as long as the drug name appears in each query.

Dependent Variable: Total Sessions

	DTCA - Stock		**Components - Stock**	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	59.19***	23.10	155.30***	27.82
age	0.26	0.57	1.20**	0.56
dtca	5.17***	0.75	-	-
alltv	-	-	2.71***	0.75
allmags	-	-	0.85	0.71
allnewsp	-	-	-0.26	0.77
allradio	-	-	2.15**	0.98
outdoor	-	-	1.51	1.42
internet	-	-	3.74***	0.80
rld	-11.70*	8.33	-16.06**	7.86
observations	510		510	
R^2	0.21		0.31	

*Notes: ***, **, * Significant at the 1%, 5%, and 10% level respectively. Drug class fixed effects included but not shown. All advertising variables are in logs.*

Table 2.7: Regression Results - Frequency of Search

The two columns of the table each contain a different breakdown of DTCA. Specification one includes the cumulative stock of advertising on a drug from January 1994 through February 2006, just prior to the time-frame of the AOL click-through data. The second specification includes a breakdown of spending by media type. I

attempt to limit the endogeneity that may exist between DTCA and search activity by including only DTCA spending prior to the period I observe the search sessions. In addition, most DTCA is offline with only 3% of total DTCA in the form of online spending.[15]

Overall DTCA is positive and significant meaning that increased ad spending leads to an increase in the number of search sessions performed on a drug. Focusing on the breakdowns by media category reveal that television, internet, and radio have positive and significant effects, consistent with the notion that these ads provide relatively less detailed information and may leave a consumer wanting to seek out additional sources. Spending on outdoor advertisements is insignificant, though the result may be misleading due to this category being the smallest of the types. Also as expected, newspaper and magazine spending is largely insignificant which may imply that consumers are able to find all the information they need in these ads.

The drug's age since original FDA approval is positive in both regressions (and significant in the second), which is somewhat surprising, but it may be driven by a few older but very popular drugs. Finally, a drug that is the innovator or pioneering version of a medicine reduces the number of search sessions and drug class fixed effects (not shown in the table) are largely insignificant, though central nervous system drugs and psychotherapeutic agents are searched upon relatively more frequently.

In Jin and Iizuka (2005), they find that the effect of a drug's DTCA on the propensity of consumers to visit their doctor regarding that drug, depreciates by only

[15]Online spending on paid-search advertising is larger, though not included in the data.

about 4% per month. However, in Jin and Iizuka (2007), they find that the effect of DTCA on the likelihood that a doctor prescribes a drug is small and depreciates almost immediately. In table 2.8 I present the results of two additional specifications which assess the rate at which DTCA spending depreciates in terms of its influence on search activity.

Dependent Variable: Total Sessions. Depreciation Analysis

	DTCA - Quarters		Components - Quarters	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	115.30***	23.26	431.57***	118.21
age	1.77***	0.55	1.74***	0.52
dtca_1_qtrb4	3.15***	1.16	-	-
dtca_2_qtrb4	3.29***	1.39	-	-
dtca_3_qtrb4	0.76	1.35	-	-
dtca_4_qtrb4	0.65	1.09	-	-
alltv_1_qtrb4	-	-	8.04***	2.22
alltv_2_qtrb4	-	-	-6.54**	3.01
alltv_3_qtrb4	-	-	-1.37	3.10
alltv_4_qtrb4	-	-	2.64	2.30
allmags_1_qtrb4	-	-	1.25	1.24
allmags_2_qtrb4	-	-	2.02*	1.47
allmags_3_qtrb4	-	-	-0.38	1.41
allmags_4_qtrb4	-	-	-1.43	1.17
allnewsp_1_qtrb4	-	-	8.57***	2.63
allnewsp_2_qtrb4	-	-	0.60	2.18
allnewsp_3_qtrb4	-	-	-3.77*	2.34
allnewsp_4_qtrb4	-	-	2.61	2.61
allradio_1_qtrb4	-	-	-1.34	3.87
allradio_2_qtrb4	-	-	-4.48	3.69
allradio_3_qtrb4	-	-	10.24**	4.47
allradio_4_qtrb4	-	-	-2.53	3.97
outdoor_1_qtrb4	-	-	15.84***	4.40
outdoor_2_qtrb4	-	-	-8.74	8.10
outdoor_3_qtrb4	-	-	-9.17**	4.58
outdoor_4_qtrb4	-	-	12.42**	5.63
internet_1_qtrb4	-	-	3.30**	1.73
internet_2_qtrb4	-	-	0.04	2.07
internet_3_qtrb4	-	-	0.13	2.32
internet_4_qtrb4	-	-	2.63*	1.99
rld	-22.50***	7.94	-18.15***	7.54
observations	510		510	
R^2	0.30		0.43	

*Notes: ***, **, * Significant at the 1%, 5%, and 10% level respectively. Drug class fixed effects included but not shown. All advertising variables are in logs.*

Table 2.8: Regression Results - Depreciation Analysis

The first regression specification includes overall DTCA separately for the

last four quarters and the second displays the results of a similar regression on the components of DTCA. The results show that after two quarters (or six months), the positive and significant effect of DTCA disappears. However, in the regression on the advertising components, I see that although television and internet advertising are very effective in the most recent quarter, the effect is zero or even negative for quarters two through four. While magazine spending remains insignificant, spending on newspapers in the most recent quarter is strongly positive and significant and then fades for less recent spending.

2.4.2 Depth Regressions

In addition to investigating the influence of DTCA on search frequency, I also present evidence on how advertising affects the *intensity* of search. Consumers who are exposed to a drug advertisement on television may go to a search engine for additional details about the drug or information relating to price and purchase availability. This information may come from multiple sources such as the pharmaceutical companies (e.g., pfizer.com), government sites (e.g., FDA.gov), and advertising-driven medical information sites (e.g., webmd.com). Different forms of offline advertising may affect how intensively a consumer searches these sites.

I analyzed several measures of search intensity including the number of clicks in a search session, the length of a session in minutes, and the number of query revisions (drill downs) preformed. One important measure, which is reported here, is the likelihood that a search session goes beyond the first page of results. Since

drug information tends to be complicated and has many dimensions, consumers may be more prone to search deep into the results to find accurate and unbiased information about a drug, especially following an advertisement that provides very little detailed information. Table 2.9 displays the results of a probit regression modeling the probability that a user clicks on a result beyond the first page in a search session.

Dependent Variable = 1 if user clicked beyond page 1 of the search results

		DTCA - Prev. Quarter		Components - Prev. Quarter	
Parameter	**Mean**	**dY/dX**	**z-stat**	**dY/dX**	**z-stat**
age	10.000	0.0006***	2.9482	0.0008***	4.2705
dtca	-10.754	0.0005***	2.5683	-	-
alltv	-12.711	-	-	-0.0008***	-3.1519
allmags	-11.397	-	-	-0.0002	-0.7384
allnewsp	-13.419	-	-	0.0018***	5.6919
allradio	-13.561	-	-	0.0022***	5.2635
outdoor	-13.777	-	-	-0.0294	-0.0765
internet	-12.278	-	-	0.0006**	1.7678
rld	0.481	0.0016	0.5704	0.0016	0.6100
observations		28,679		28,679	
percent concordant		52.1		55.3	

*Notes: ***, **, * Significant at the 1%, 5%, and 10% level respectively. Drug class fixed effects included but not shown. All advertising variables are in logs.*

Table 2.9: Regression Results - Depth of Search

Similar to the regressions in the last section, I report two specifications including the effects of overall DTCA and its components, focusing on the quarter immediately prior to the search sessions. I report the mean of the variables and as well as their marginal effects (i.e., the predicted change in the probability for a one unit change in the independent variable at the mean). The z-statistics are also reported.[16]

[16]Note that to measure the fit of the model, I report the *percent concordant*, which is the percent of observation pairs such that the observation with the higher ordered response corresponds to the higher predicted response. In my sample, only about 5% of sessions involve clicks beyond page one, so the dependent variable in my regression is unbalanced and the predicted probabilities are skewed towards zero. Calculating a pseudo-R^2 by defining a correct prediction of success as a

Overall DTCA has a positive and significant effect on the likelihood of a more intense search session. Considering the breakdown by media category in the second specification, positive and significant effects are found for newspaper, radio and internet ads, which is the same result I found in the regressions on search frequency. The negative effect of television ads is consistent with the notion that television ads refer a consumer to the drug's website for more information and this site often appears high in the ranks. Therefore a consumer simply using the search engine as a navigational tool to reach a predetermined page (e.g. lipitor.com) instead of typing in the URL directly, will have a less intense search session.

2.5 Conclusion

The analysis has shown that consumers seek diverse information about prescription drugs online and their behavior is influenced by the online and offline advertising to which they are exposed. Offline advertising not only increases the likelihood that a user searches for a drug, but also increases the depth of search within a search session. Consumers searching for drug information also behave differently than those seeking information about consumer products like electronics. Overall, drug sessions tend to feature more clicks on different search results and these clicks come faster than in electronics sessions. It may be that consumers are seeking specific information about a drug and can quickly determine if a search result is going to provide it.

predicted probability above 0.5, as Greene and others suggest, would result in a very high measure of fit, but only because most of the observed and predicted outcomes are zero.

Among the drug searches, activity is evenly spread among younger and older drugs. Advertising spending on those drugs is slightly skewed toward younger drugs though there is still significant spending on drugs that are 8-10 years old. Click patterns within a search session reveal, as expected, more clicks on dot-gov and dot-org/net/info results and the popularity of these sites grows in longer sessions. Consumers may be immediately clicking on the first or second result (usually a dot-com) but then will make a transition away to results with other extensions, perhaps in an effort to seek unbiased information. The distribution of clicks by search result rank also reveals that consumers are more likely to click on lower ranked results further down the results page.

In the regression analysis, I analyzed the effects of DTCA on search frequency and depth. Overall DTCA increases both the frequency and depth of search, though the various types of DTCA (via different media), each affect search differently. DTCA that provides only a major statement regarding a drug and few additional details such as television and internet banner ads, increase the frequency of search. This may be the result of the FDA regulation stating that these ads must direct consumers to seek additional details at the drug company's website. If they are simply using the search engine to find this site, their search session will likely be very short and the evidence suggests this effect with a significant and negative coefficient on television DTCA in the depth regression.

Finally, the analysis of depreciation shows that the effects of DTCA spending disappear after about six months. Television and internet advertising have a strong positive effect on search in the near term, though quickly fades even after just three

months.

Moving forward, I plan to focus on the effects of television ads, the largest class of DTCA, on search activity using a detailed dataset from TNS which includes the exact time and placement of an ad during a broadcast. With the growing accessibility of laptop computers including netbooks, consumers are likely reacting quickly to television advertisements and immediately seeking further information on the internet. Combining this with either the AOL dataset analyzed here, or a new dataset from comScore which also tracks household internet use, I can determine the effects of television DTCA, including how varying demographics influence the effects of an ad on consumer search behavior.

Chapter 3

Drug Information via Online Search Engines

3.1 Introduction

Search engines are the gateway to the internet as 94% of internet users access engines to find information on the web.[1] According to Nielsen Rankings, over 9.5 billion searches were executed on the top 10 search engines in the US in March of 2009, 16.7% higher than the year before. The five largest engines by number of searches are Google (64.2%), Yahoo (15.8%), MSN (10.3%), AOL (3.7%), and Ask (2.1%), with Google driving most of the growth in search.[2] The availability of health care and drug information on the internet is arguably one of the more important areas in need of study given the important public health consequences. In this paper, I document the supply and content of this type of information on four large search engines and across time.

Given the vast amount of information on the internet, one could study the supply of search results related to many different industries, though I focus on the prescription drug market. A 2008 Nielsen study found that health websites are consumers' second most important source of medical information behind their doctor. About 50% of the US internet population visited a health-related website in July of 2008. In the Nielsen study, 82.6% of subjects reported having visited a

[1] See Ghose and Yang (2008).
[2] See `www.nielsen-online.com`.

website for health information at some time in the past, and a third of those used a search engine to find the information they were seeking. Overall, drug queries involve the potential for users seeking a wide variety of complicated information, so the summary text and the source (domain and extension) of a search result will likely be important determinants of a user's attention and click behavior.

A complication that one faces when studying the supply and demand of information via a search engine is that it is a very dynamic market that is constantly evolving. The supply (search results) influence the demand (consumer search behavior) and vice versa by way of the engine's ranking algorithm, and this creates an endogeneity problem for the analysis. Since the algorithms are proprietary, it is impossible to know how much, for example, the rank of a search result is purely a function of its relevance to the search query versus a function of the attention garnered from being of a certain rank in the past. One way to mitigate this problem is to average certain metrics across time, which I do frequently in the analysis.

There are two types of search results that appear on a search engine when a user submits a query. Organic results are those generated by the engine's algorithm as being the most *relevant* to the user's query. Relevance is determined differently by each engine and may include determinants such as past click traffic and the number of inbound links to a site from other relevant websites. The title and summary text appearing on the search engine is determined endogenously by the engine itself.

Sponsored or paid results are those that appear (at times) above, below, and to the right of the organic results. See Athey and Ellison (working paper) and Varian (2007) for details on the auction mechanism and optimal bidding strategies for

sponsored results.[3] Their placement is driven both by relevance and by the amount that the advertiser has paid to be listed. The title and summary text is chosen by the advertiser. It is often difficult to distinguish between organic and sponsored results, undoubtedly because the search engine generates revenue from them only when a user clicks on a sponsored result.[4] I will analyze the different content and domain extensions between the two types of results, though it is clear that sponsored results tend to be more promotionally driven and, for drug searches, dominated by online pharmacies. Ghose and Yang (working paper) analyze the substitution pattern between organic and sponsored links for a specific website address, or Uniform Resource Locator (URL), and generally find that there are positive and asymmetric spillovers from one type of link to the other.[5]

I consider four large search engines: Ask, Google, MSN, and Yahoo. I do not include AOL's search engine, which has a similar market share to Ask, though through a partnership with Google, AOL uses Google's algorithm to generate both their organic and sponsored links.[6] Ask also partners with Google to display their sponsored links in addition to Ask's self-generated links.

In the analysis, I show the popularity of different website extensions, also called *top-level domains*, such as dot-com and dot-gov. Consumers may choose to click relatively more frequently on, for example, a dot-gov site in order to find accurate and unbiased information, knowing that only the US government can register a

[3]See also: Edelman, et. al. (2007) and Ghose and Yang (2008).

[4]Sponsored results often appear with a slightly different background than the organic results and in my experience, it is increasingly difficult to tell them apart.

[5]In future work, I hope to extend this type of substitution analysis to drug-queries.

[6]See `http://www.nytimes.com/2007/04/09/technology/09iht-aol.1.5197096.html?_r=1`.

website with a dot-gov extension.[7] These extensions are maintained by the Internet Assigned Numbers Authority (IANA), who regulate which sites can have an address ending in each extension.[8]

The remainder of this paper is organized as follows. Section 2 provides a description of the data including the list of drugs I use and a method for generating the content of each search result. A descriptive and regression analysis of the data is developed in section 3, including the differences in supply and content across engines and the dynamics of a URL's rank over time. Section 4 concludes and provides some directions for future work.

3.2 Data

Drug Selection

To select the list of queries, I started with the 2004 National Ambulatory Medical Care Survey (NAMCS), and determined the 20 most popular drug classes based on "drug visits" which is the number of visits to a doctor in which a given drug is prescribed.[9] Of these, I decided to focus on the top 95% of drugs in three National Drug Code (NDC) classes: antidepressants, cholesterol, and diabetes, due to their relatively high advertising intensity online and offline. Since NAMCS only contained drugs approved through 2004, I supplemented the list with recently approved drugs

[7] See Huh and Cude (2004), which analyzes medical-related websites to calculate a measure of bias based on the type of information appearing on each page.

[8] A complete list of top-level domains and their requirements can be found at: `http://www.iana.org/domains/root/db/`.

[9] See `http://www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm`.

in each of the three classes from FDA's Orange Book.[10] Starting in 2006, NAMCS started using a different coding system for all drugs. Each drug can belong to up to four categories which sometimes span the classes from the old NDC system. I use the old class codes in this paper as they are broadly in-line with the new system.

This yielded 99 unique brand names that formed the basic search list. I supplemented these queries in several ways. First I paired the top five drugs in each class (based on total search results) with each other to assess queries where a consumer was seeking comparison information about two similar drugs. I also added keywords to the top five drugs in each class where the keywords where determined using Google's Adwords tool.[11] These include risk-related keywords like "interactions" and "side effects" as well as sales promotion keywords like "discount" and "price." Finally, for brand name comparisons and brand names paired with keywords, I included searches with and without quotes. In all, this yielded 458 search queries.[12]

Crawler Data

With the help of two excellent research assistants,[13] we designed a web crawler that submitted the list of 458 search queries to four large search engines (Ask, Google, MSN, and Yahoo) every day at 12:00pm during the period from February - September 2007. The crawler saved the top 100 organic search results which appeared on first 10 pages. These first 10 pages also contain sponsored search

[10] See `http://www.fda.gov/cder/orange/obreadme.htm`.
[11] See `https://adwords.google.com/select/KeywordToolExternal`.
[12] See the appendix for the complete list.
[13] Chien (Daniel) Yin and Chris Wasko.

results, the number of which varies depending on the query.[14]

Since the crawler program returned the raw HTML files containing the search results for each engine-day-query, we then wrote a parsing program to separate out the following fields/variables for each result: rank, title text, summary text, URL (displayed and actual)[15], result type (organic or sponsored), and result position (for sponsored results).

Basic Statistics	
Organic Results	
Ask	4,200,829
Google	10,604,000
MSN	10,724,339
Yahoo	10,725,091
Sponsored Results	
Ask	3,384,565
Google	2,667,023
MSN	3,949,351
Yahoo	6,364,854
Date Range	Feb - Sep 30, 2007*
Unique Queries	458
Query Types	
Drug Name Only	99
Drug + Informational Keyword	195
Drug + Promotional Keyword	96
Drug + Drug	68
Drug Classes	
Depression	161
Cholesterol	133
Diabetes	164

**Data from the Ask search engine is only available through May and does not include the organic links ranked 91-100.*

Table 3.1: Basic Statistics

Table 3.1 displays the basic statistics of the data collected by the crawler. Due to a parsing error, only a limited sample was gathered from the Ask search engine.

[14]We faced several challenges in collecting the data including adapting to formatting changes on each engine that occurred during the time period and adding a random time increment between queries to avoid the search engine (correctly) flagging us as a crawler. We assume our own search activity has minimal impact on the supply of search results.

[15]These are frequently different especially for sponsored results which are routed through the search engine first (so the engine can charge the advertiser) before taking the user to their destination page.

Classification Algorithm

In order to determine the type of search results that were appearing following each query, I devised an algorithm to classify each search result as being either informational, promotional, or neutral. This was accomplished with the following steps:

1. For all 4 engines and for one week, first collect all words appearing in the top 100 organic and all sponsored search results following two types of queries:

   - drug name + "buy" or drug name + "cheap" (likely promotional sites)
   - drug name + "information" or drug name + "side effects" (likely informational sites)

   where drug name was one of the 99 brand names in the sample. I do this separately for titles and summaries and for organic and sponsored results, which provides 8 lists of words.

2. Create a frequency table of all of the words appearing in each list and save the top 200 most popular words in each list.

3. Eliminate any words that appear in both categories (informational and promotional) and save the top 50 unique words in each category.[16]

4. Analyze every search result in the database and calculate the proportion of words in each text field that also appear in the corresponding top 50 list. E.g., an organic summary text field may have 25% promotional words and 10% informational words.

[16]The uniqueness requirement also eliminates common words that frequently appear in text fields, but are unhelpful in classifying content.

With these proportions in hand, I can form a metric called the "average content" of a search result which is simply the difference between the proportion of words that are promotional and the proportion that are informational. I can also create a binary indicator of content and, for example, classify a result as promotional if it contains a relatively higher proportion of promotional keywords. The keywords used in the classification are shown in table C.2 in the appendix. Note that some of the words are actually numbers, which are very common in promotional results, and therefore helpful in their classification.

3.3 Descriptive Analysis

3.3.1 Supply

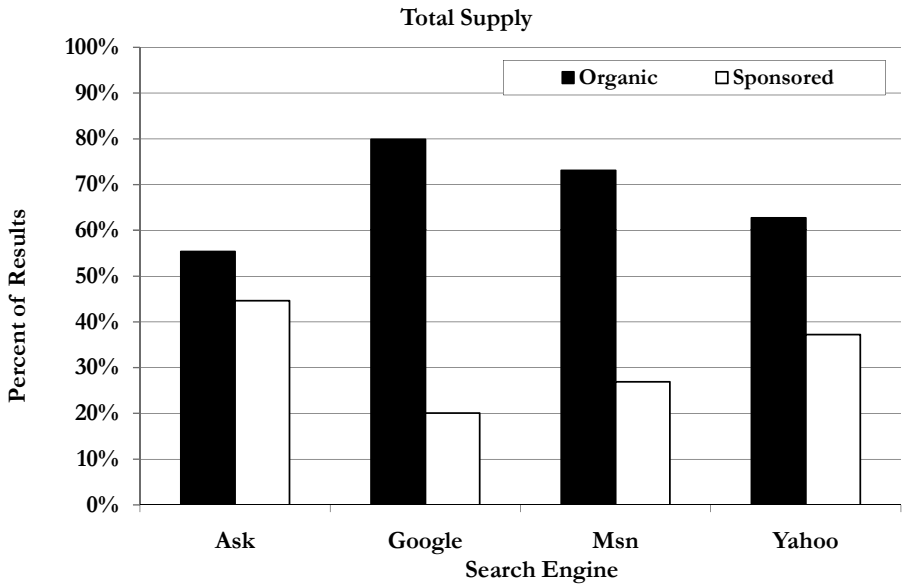


Figure 3.1: Distribution of Organic and Sponsored Results

Figure 3.1 show the overall supply of results on each engine. Note that even

with the limited sample from the Ask engine, it has relatively more sponsored results than the other engines given that it displays its own results along with those from Google. Of the other 3 engines, while they all have about the same number of organic results, Google has the largest proportion. There are usually 100 organic results collected per query-day, but for some queries there are fewer.[17]

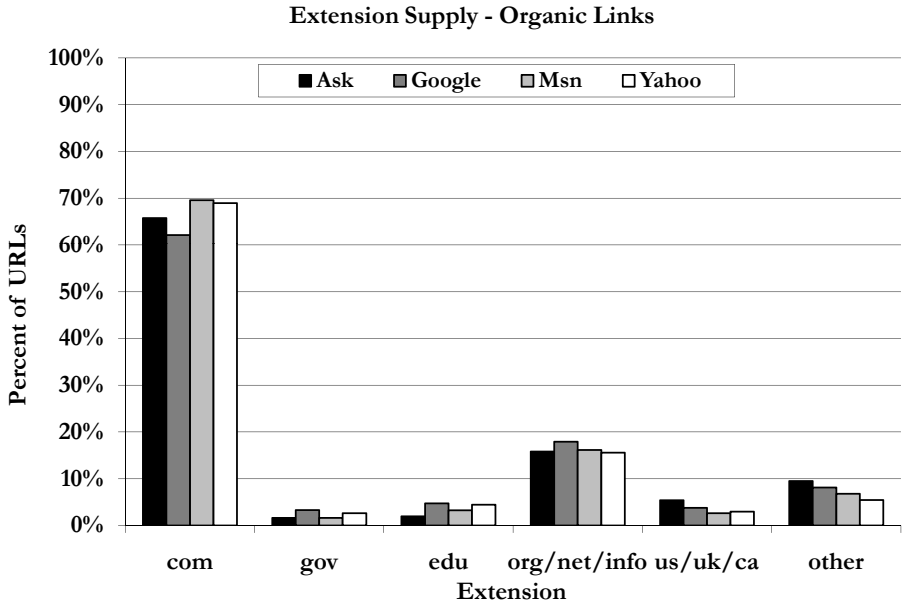


Figure 3.2: Extension Popularity - Organic Results

Organic and sponsored result popularity by extension are shown in figures 3.2 and 3.3 respectively. Google's organic results feature the fewest dot-com and the most dot-gov, dot-edu and dot-org/net/info results.[18] MSN has the largest percentage of dot-com results and fewest dot-govs. Among sponsored results, most have dot-com extensions, except for MSN who has relatively more dot-org/net/info

[17]In theory, with 235 days and 458 search queries, I could observe a maximum of 235*458*100 = 10,763,000 observations per engine. For Google, MSN, and Yahoo, I observe 99% of this theoretical maximum.

[18]The differences between the engines are largely statistically significant. For example, Google's percentage of dot-gov results is statistically higher than each of the other three engines.

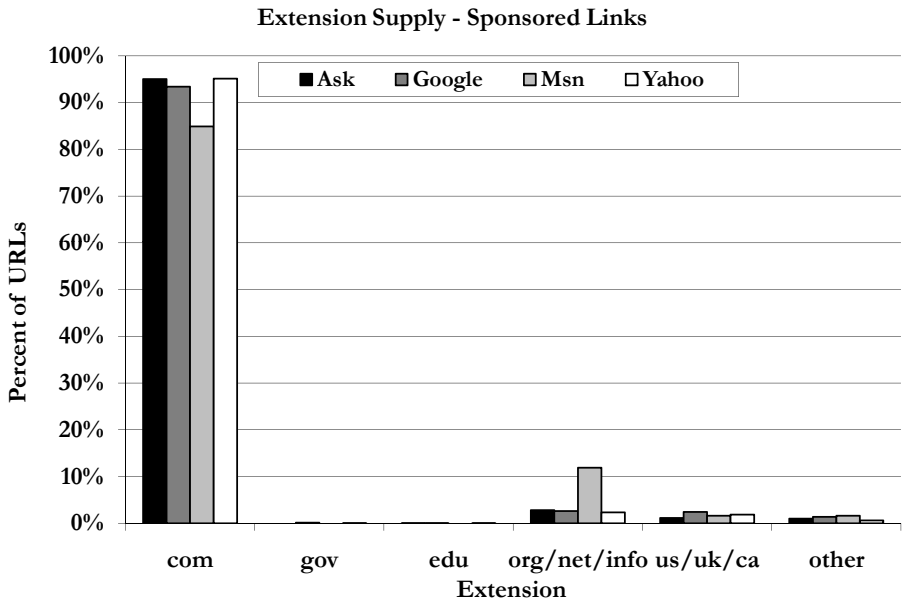


Figure 3.3: Extension Popularity - Sponsored Results

extensions among their sponsored results.

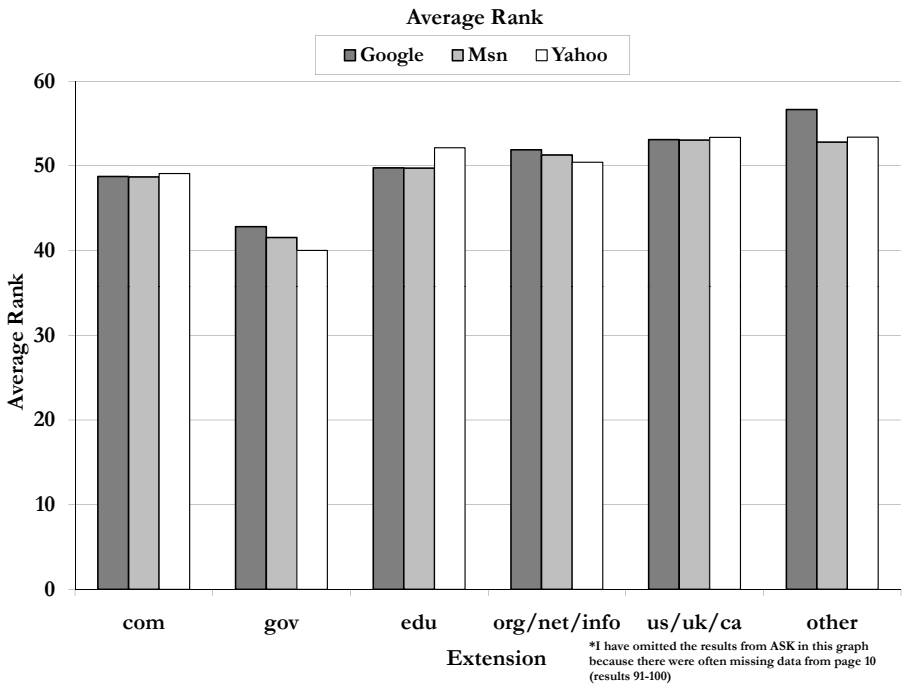


Figure 3.4: Average Rank by Extension - Organic Results

I finally break down the average rank of organic results by extension in figure

3.4. I omit Ask because of the parser problem. If the results were spread evenly, they should have a mean of 50, but here dot-gov sites tend to be pushed toward the top of the page (lower numbered ranks). Of the three engines, the dot-gov sites on Yahoo are most likely to appear high in the search results.

3.3.2 Content

Using the rank popularity from the AOL click-through database (among all queries), I calculate an *attention index* for each organic rank because links appearing toward the top of the results are more likely to receive a click than those lower in the results. The index is simply the proportion of clicks on each organic rank, from 1 to 100.[19] Then I calculate the percent of organic results weighted by the attention index for which their summaries are classified as promotional, information, or neutral. E.g., a result is promotional if it contains a higher proportion of promotional keywords compared with informational.

Figure 3.5 displays the attention weighted content of each engine. MSN's results tend to be more promotional than other engines and Google's are more informational. Classification reflecting the actual proportions are reported in the kernel density figures. Figure 3.6 is the same breakdown for sponsored results. Here, Google and Yahoo tend to be relatively more promotional and Ask and MSN are more informational.

Figures 3.7 and 3.8 display the organic and sponsored summary content bro-

[19]For example, because users click on the first result much more often than other results, the first rank receives a weight of 0.423 while the fifth rank has a weight of 0.049.

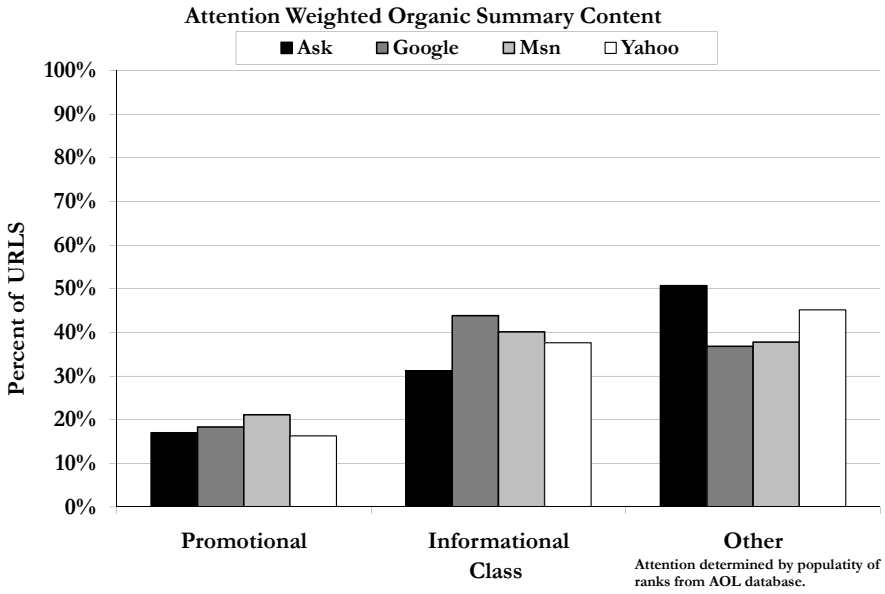


Figure 3.5: Content of Summary Field - Organic Results

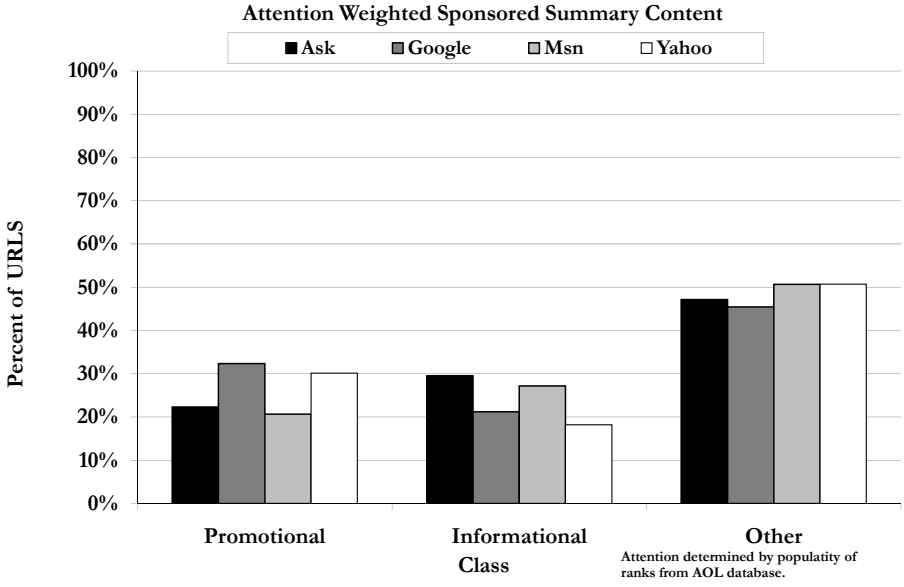


Figure 3.6: Content of Summary Field - Sponsored Results

ken down by extension. For organic results, dot-com and dot-gov sites are more informational, while surprisingly, dot-edus are relatively more promotional for all engines. Further investigation revealed that, e.g., the engines are picking up on comments left on university bulletin boards by online pharmacies trying to sell their

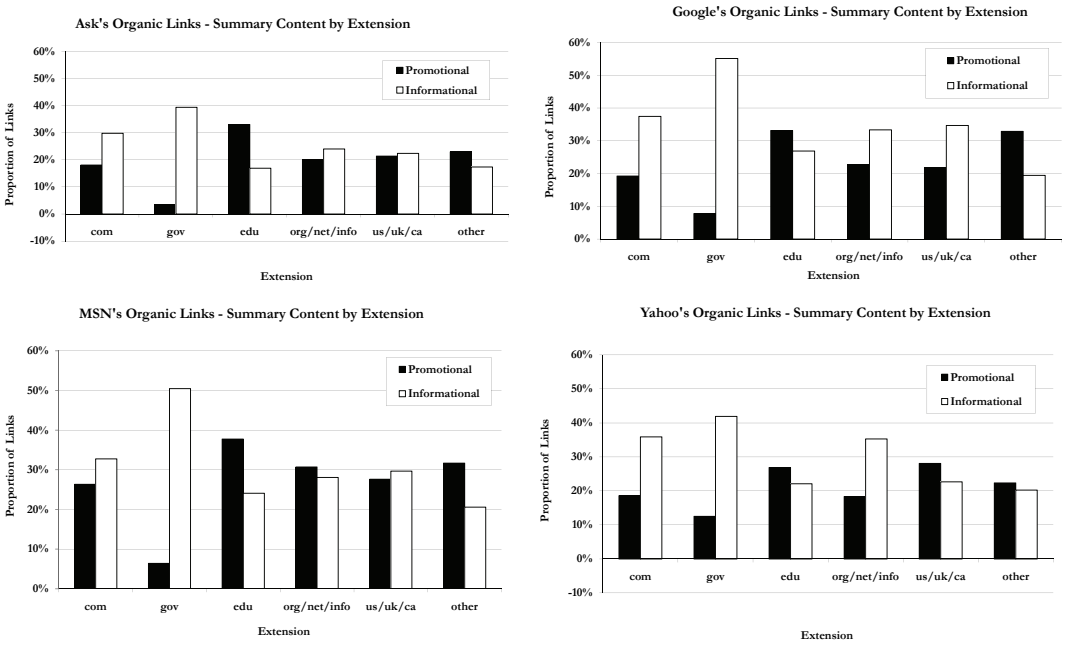


Figure 3.7: Content of Summary Field - Organic Results - By Extension

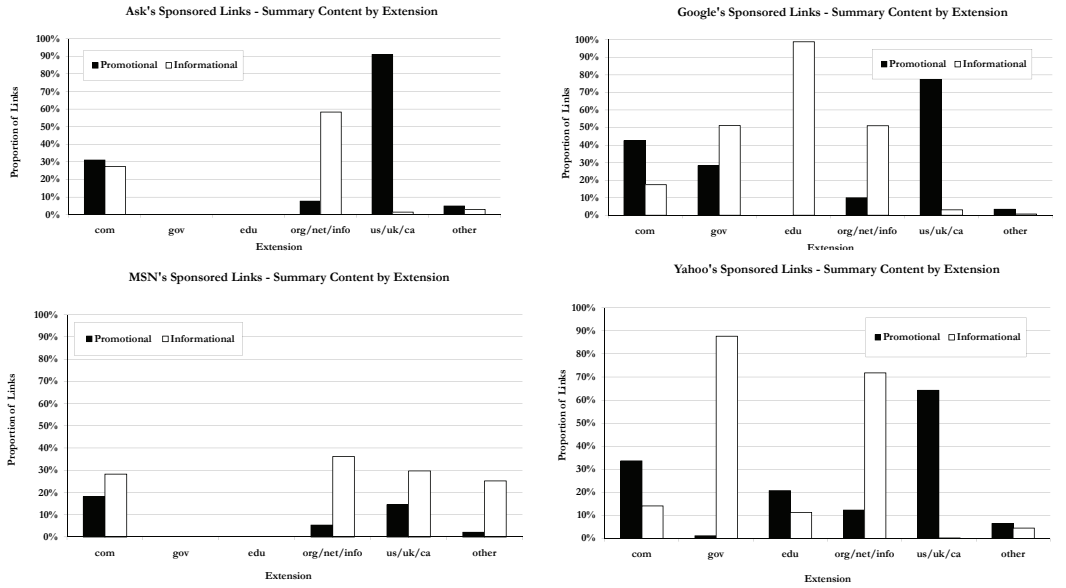


Figure 3.8: Content of Summary Field - Sponsored Results - By Extension

drugs. As for sponsored results, Ask, Google and Yahoo feature mostly dot-coms and these tend to be relatively promotional as expected. Sponsored results ending

in dot-org/net/info tend to be informational in content. MSN is unique in that all of its sponsored results tends to be more informational, in line with the large proportion of their results ending in dot-org/net/info.

3.3.3 Rank and Content Comparisons

An additional approach to comparing the results from a query across search engines is to analyze the rank and contents of a set of organic results. In figures 3.9 and 3.10, I display a comparison of Google and Yahoo. The first scatter shows the ranks of identical URLS (following the same query on the same day). The differences in algorithms is clear from the figure and a weak correlation of 0.37. However, when comparing the proportion of promotional keywords on the same two engines, a stronger correlation (0.44) is revealed. Thus, though the algorithms differ in how they rank the results, the process for selecting which words and phrases to include in the summary text is similar. Repeating this for other engine comparisons reveals roughly the same pattern though the correlations are less strong.

Though the rank of a search result may be very different on a given day, there may be some relationship between the changes in the rank over time. In figure 3.11, I track the rank of the same URL (following the same query) across time for the three engines for which I have complete data. Here I see that the ranks on MSN and Google are fairly stable though there are frequent spikes in Yahoo's rank. These may be due to algorithm testing by the engine throughout the year.[20]

[20]In the future, I will analyze how exogenous shocks, such as a FDA news story about a drug, affect the rank dynamics of specific URLs or extension classes across search engines.

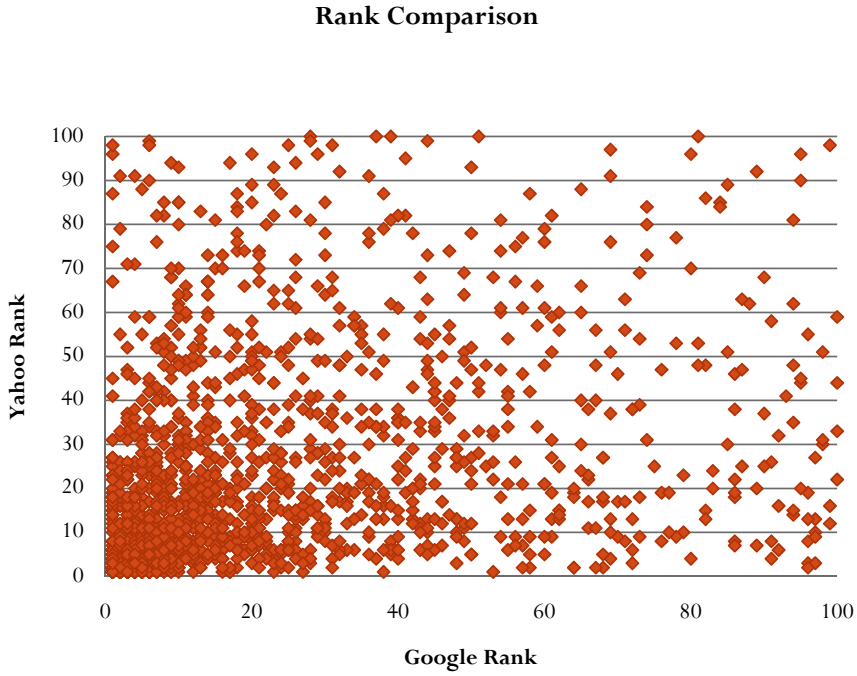


Figure 3.9: Rank Comparison - Organic Results - Google vs Yahoo

3.3.4 Kernel Density Plots of Content

As a final analysis of the content differences between search engines and across different extensions and result types, I estimate Gaussian kernel density distributions using the difference between the proportions of promotional and informational keywords in each result. I first drop the search results that have no promotional or informational keywords (i.e., those that would be classified as neutral/other). The variable plotted (PropPromo - PropInfo) ranges from -1 to +1 with -1 corresponding to a result that is completely informational and +1 meaning the result was completely promotional. A value of zero means that a result contained an equal (non-zero) number of informational and promotional keywords.

Figure 3.12 shows that organic results on all engines tend to be more infor-

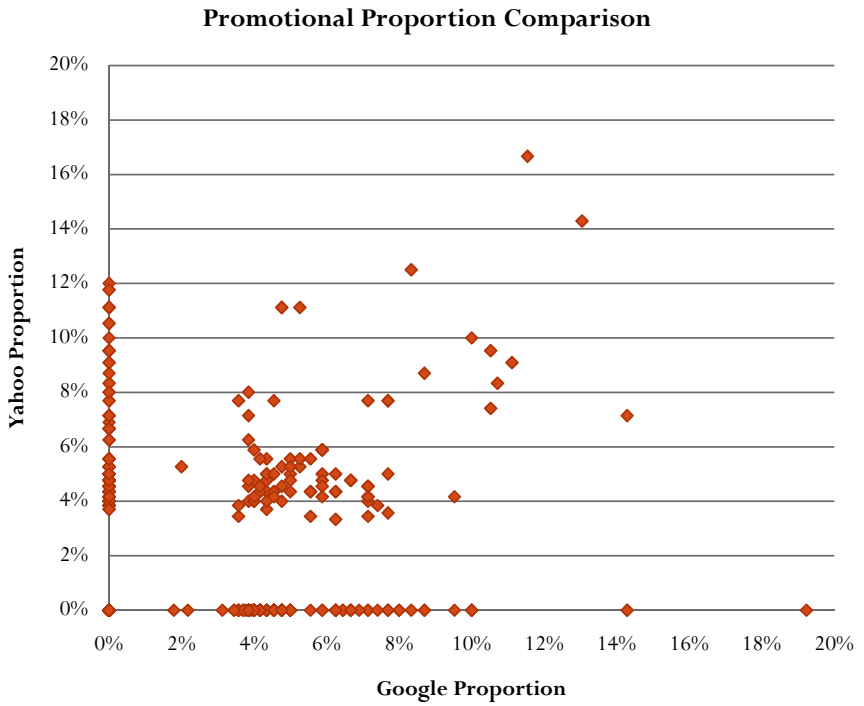


Figure 3.10: Summary Content Comparison - Organic Results - Google vs Yahoo

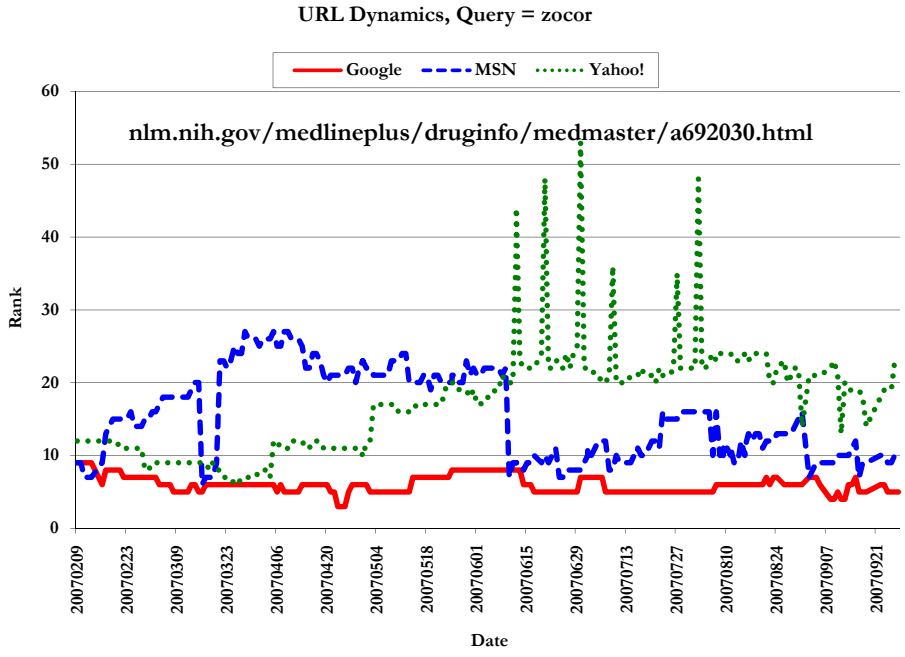


Figure 3.11: Organic Rank Dynamics

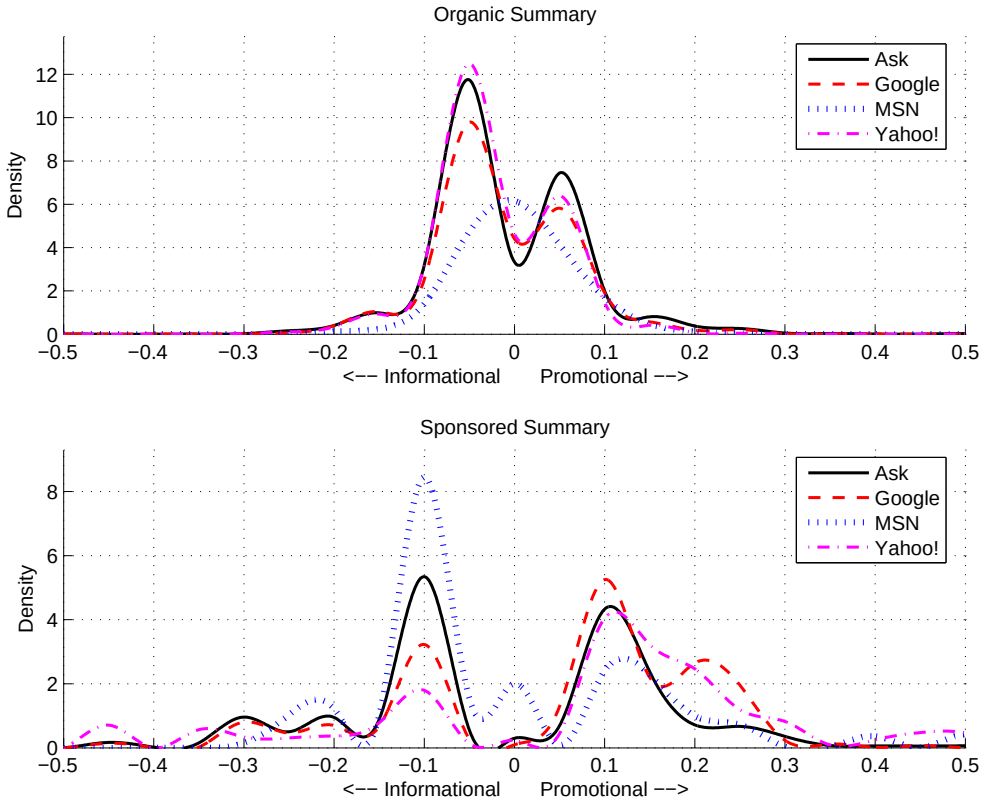


Figure 3.12: Kernel Density of Summary Content

mational with the spike around -0.05, though MSN has the highest density of more promotional sites. Sponsored results tends to be either very information or very promotional, as revealed by the heavy tails in each distribution. MSN tends to have the most results with informational sponsored results (consistent with the large percentage of dot-org/net/info sites in their sponsored results).

In the appendix, I display a breakdown of summary content of organic and sponsored results by extension (see figures C.1 and C.2). For organic results, dot-com and dot-gov results are again shown to be more informational for all engines. Dot-edu sites display about just as many promotional as informational sites, with the heaviest tail for promotional content on Google. For dot-org/net/info, most

of the results are informational as expected. Among the sponsored results, dot-com sites again account for most of the sponsored results and tend to be either very informational or very promotional. There are very few dot-gov and dot-edu results among the sponsored results and the dot-org/net/info sites tend to be very informational.

3.3.5 Probit Analysis

Finally, I report the results of a simple probit regression analyzing the determinants of rank in a search engine's results. While I could consider the likelihood that a URL achieves a given rank using an ordered probit approach, since most users do not venture beyond the first page of search results, I only consider the probability that a result appears on the first page. In this regression, I consider organic results from March 2007 on all 4 engines in the sample. Since characteristics of individual drugs (like drug age and advertising intensity) do not vary across the ranks in the search results, I cannot analyze the influence of these variables on where a URL appears. However, I can interact them with the extension of a search result's URL since they do vary by rank. I can then determine, for example, how the age of a drug may affect the likelihood of a dot-gov URL appearing on page one of the search results.

Table 3.2 displays the result of a probit estimation on the probability that a result appears on page one as a function of website extensions, extension/age and extension/advertising interactions, and result summary contents. Definitions for

Dependent Variable: Pr(Page1)								
	Ask		Google		MSN		Yahoo	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	-1.711***	0.023	-2.220***	0.039	-1.704***	0.032	-1.730***	0.023
dotcom	0.492***	0.025	0.670***	0.040	0.312***	0.033	0.220***	0.025
dotgov	1.049***	0.049	0.810***	0.051	0.732***	0.053	0.804***	0.041
dotedu	0.357***	0.110	-0.481***	0.107	0.127	0.158	-0.141**	0.069
dotorgnetinfo	0.465***	0.033	0.553***	0.046	0.222***	0.040	0.247***	0.032
dotintl	0.289***	0.030	0.069*	0.049	-0.033	0.046	-0.121***	0.035
dotcom_age	0.001	0.001	0.002**	0.001	0.007***	0.001	0.003***	0.001
dotgov_age	-0.016***	0.003	0.006***	0.002	-0.010***	0.003	0.001	0.003
dotedu_age	0.021***	0.006	0.067***	0.005	-0.034***	0.011	0.035***	0.004
dotorg_age	-0.007***	0.002	0.000	0.002	0.010***	0.002	0.010***	0.002
dotcom_dtc	-0.222***	0.030	-0.220***	0.030	0.098***	0.025	-0.060**	0.029
dotgov_dtc	0.530***	0.104	0.379***	0.095	-1.284***	0.211	-0.181**	0.099
dotedu_dtc	-2.036***	0.500	-0.219	0.173	0.870***	0.331	-1.625***	0.249
dotorg_dtc	0.215***	0.064	0.675***	0.058	-0.663***	0.074	0.120**	0.061
prop_promo_summary	-0.730***	0.135	-0.945***	0.155	-2.393***	0.145	-0.163	0.143
prop_info_summary	1.468***	0.086	5.894***	0.075	4.500***	0.105	4.642***	0.083
prop_promo_title	1.893***	0.086	-2.113***	0.153	-2.220***	0.133	0.091	0.086
prop_info_title	0.771***	0.039	1.467***	0.031	1.524***	0.046	0.921***	0.034
observations	148,718		176,700		176,431		176,680	
percent concordant	61.6		72.5		65.9		65.7	

Notes: ***, **, * Significant at the 1%, 5%, and 10% level respectively. Omitted categories: extension = other, summary content = other, title content = other, drug class = diabetes. Organic results from March 2007.

Table 3.2: Regression Results: Probit of Pr(Page 1)

all variables used in the regression are summarized in table C.3 in the appendix. Promotional sites are uniformly pushed down and informational sites are more likely to appear on page one. Dot-gov sites are the most likely of all extensions to appear on page one with the greatest effect for Google's engine. In all but Ask's engine, dot-edu and international sites tend to get pushed off of page one. The interaction terms reveal that, for most engines, older drugs are more likely to have dot-com sites appearing on page one. The reason may be that younger drugs have few promotional dot-com sites appearing high in the ranks.[21]

I am unable to perform a similar analysis predicting the probability that a sponsored link appears on page one because the dataset does not include the page

[21]Note that to assess the fit of the model, I report the *percent concordant*, as explained in chapter 2.

on which a sponsored link appears and there are often a different number of sponsored links on each page. However, since I do observe the rank, an ordered probit predicting sponsored links' overall rank revealed that, as expected, dot-com sites are more likely to appear high in the ranks and advertising intensity does not have a consistent effect on rank (via its interaction with the website extensions).

3.4 Conclusion

In addition to many offline sources, there is a large and diverse quantity of prescription drug information accessible online. Consumers are likely filtering this information and making their decisions about which sites to visit based on a search engine's results page, which includes the result's rank, title and summary text, classification as organic or sponsored, and the extension of the URL. I have shown that the information varies significantly across engines, over time, and between different website extensions.

The descriptive analysis shows that Ask has relatively more sponsored links compared with other engines, perhaps because of their agreement to deliver sponsored links from Google along with those generated from their own algorithm. Google's organic results feature relatively more dot-gov and dot-edu links and MSN's engine returns the most dot-com results. On all engines, dot-gov sites appear higher in the ranks compared with other extensions, because the engines' algorithms rank them higher for their relevance and/or because users frequently click these results.

I also analyze the content of the summary text in order to classify individual

results as informational, promotional or neutral. Overall, Google's results are relatively more informational and MSN's the most promotional, in line with the popular extensions on each engine. However, classifying websites solely on their extension may be misleading as I found that dot-edu sites actually tend to be more promotional. Among sponsored links, dot-com results are by far the most popular and, as expected, they tend to be relatively more promotional. Kernel density estimates confirm these results and also show that sponsored links tend to be either very informational or very promotional, as revealed by heavy tails in the distribution. Since the website owners are paying for each click on a sponsored link, they are likely trying to provide a very clear summary of what information the user will find if they click on the result.

Finally, the probit analysis revealed that informational sites are more likely to appear on page one of the results. Dot-gov sites are also relatively more likely to appear high in the results and the effect is largest for Google's engine. By including interaction terms of the drugs' ages and website extensions, I also show that younger drugs are less likely to have dot-com results high in the ranks.

In future work, I hope to track the dynamics of specific URLs (e.g., an fda.gov site) following a major news story about a drug being issued by the FDA. I expect to see a displacement of more promotional dot-com sites by the informational sites. While some analysis can be accomplished with the current dataset, other research will be possible once I have a complete picture of both the supply and demand for drug information from the same time period. I will soon have access to data from comScore's Media Metrix product which includes individual click-through behavior

for a set of consumers in the same time period as the crawler data. Matching these two data sources will allow me to investigate both the probability of a click as a function of result characteristics (e.g., rank, content, and extension) as well as determine the substitution/complementary effects of organic and sponsored links appearing in the same set of search results.

Appendix A

Chapter 1 Supplement

A.1 The Distillation Process

Since the various components of crude oil have different boiling points, a refinery's essential task is to boil the crude oil and separate it into the more valuable components. Figure A.1 displays a simplified diagram of a typical refinery's operations. The first and most important step in the refining process is called fractional distillation. The steps of fractional distillation are as follows:

1. Heat the crude oil with high pressure steam to $1,112$ degrees fahrenheit.
2. As the mixture boils, vapor forms which rises through the fractional distillation column passing through trays which have holes that allow the vapor to pass through.
3. As the vapor rises, it cools and eventually reaches its boiling point at which time it condenses on one of the trays.
4. The substances with the lowest boiling point (such as gasoline) will condense near the top of the distillation column.

While some gasoline is produced from pure distillation, refineries normally employ several downstream processes to increase the yield of high valued products by removing impurities such as sulfur. Cracking is the process of breaking down large hydrocarbons into smaller molecules through heating and/or adding a catalyst. Cracking was first used in 1913 and thus changed the problem of the refiner from

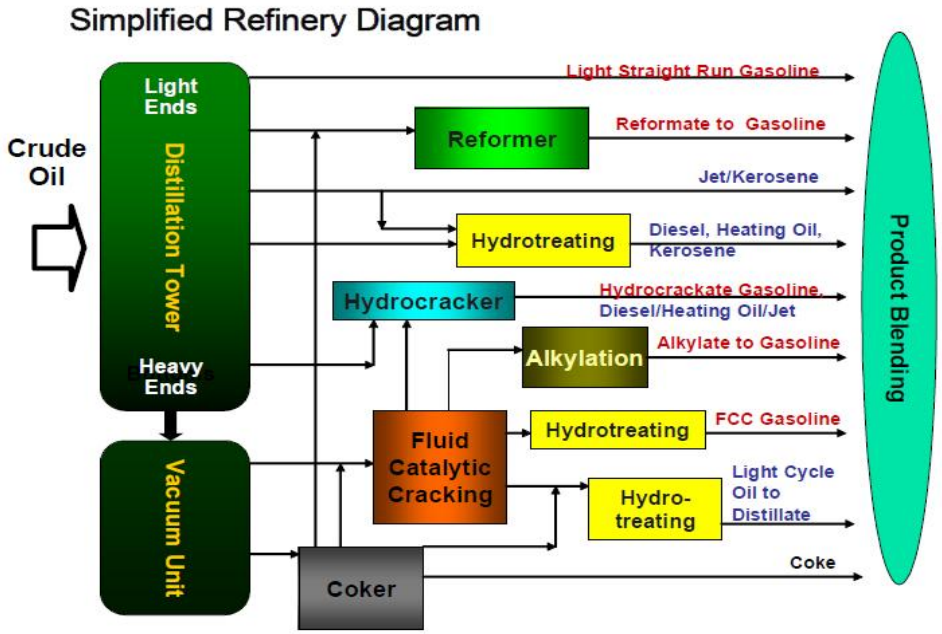


Figure A.1: Refinery Operations

choosing how much crude oil to distill into choosing an appropriate mix of products (within some range). Refineries practice two main types of cracking:

- Catalytic cracking: a medium conversion process which increases the gasoline yield to 45% (and the total yield to 104%).
- Coking/residual construction - a high conversion process which increases the gasoline yield to 55% (and the total yield 108%).

The challenge of choosing the right input and output mix given the available technology creates a massive linear programming problem.

A.2 Crude Oil Quality

Crude oil is a flammable black liquid comprised primarily of hydrocarbons and other organic compounds. The three largest oil producing countries are Saudi

Arabia, Russia and the United States.[1] Crude oil is the most important input into refineries and this raw material can vary in its ability to produce refined products like gasoline. The two main characteristics of crude that determine its quality are American Petroleum Institute (API) gravity and sulfur content. The former is a measure (on an arbitrary scale) of the density of a petroleum liquid relative to water.[2] Table A.1 summarizes these characteristics and includes some common crude types and their gasoline yield from the initial distillation process.

Table A.1: Crude Qualities

API Gravity	Sulfur Content $< 0.7\%$	Sulfur Content $> 0.7\%$
$< 22°$	Heavy Sweet	Heavy Sour - 14% yield *(Maya, Western Canadian)*
$22° - 38°$	Medium Sweet	Medium Sour - 21% yield *(Mars, Arab light)*
$> 38°$	Light Sweet - 30% yield *(WTI, Brent)*	Light Sour

Source: EIA.

Worldwide, light/sweet crude is the most expensive and accounts for 35% of consumption. Medium/sour is less expensive and accounts for 50% of consumption while heavy/sour is the least costly and accounts for 15%. Figure A.2 show how the average crude oil used by US refiners is becoming heavier and more sour over time. This means that the production costs of a gallon of gasoline are changing as refineries must invest in more sophisticated technology in order to process lower

[1]Production in this sense refers to the quantity extracted from a country's endowment.

[2]Technically, API gravity = (141.5/ specific gravity of crude at 60° F) −131.5. Water has an API gravity of 10°.

quality crude oil.

Since crude oil by itself has very little value to any industry, the price of a barrel of oil reflects the net value of the downstream products that can be created from it. The two major sources of movements in the crude oil price are upstream supply shocks (due to OPEC's quotas and hurricanes affecting oil rigs in the Gulf of Mexico) and downstream demand shocks (due to consumer's demand for refined products). The other source often sited by industry experts are refinery inventories of crude oil. Maintaining stocks of crude oil allow the refinery to respond quickly to downstream shocks like an unexpectedly cold winter increasing the demand for heating oil.

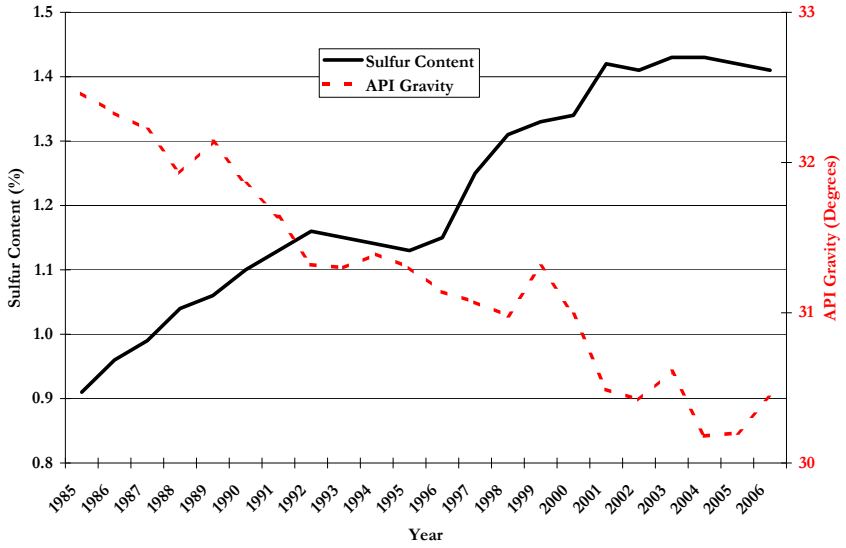


Figure A.2: Average Crude Oil Quality: Heavier and More Sour

Within the various types of crude oil, the prices of each quality respond differently to shocks. The "light/heavy" differential is one measure that indicates the benefit a refiner can achieve by investing in sophisticated equipment to process

heavier crude oil into highly-valued refined products. The differential has varied significantly over the last 10 years from 3 dollars per barrel to almost 20 dollars per barrel. An oil refinery faces a unique decision when making its production choice, one that provides for both flexibility and complexity. One one hand, consumers do not care about the type of crude oil, oxygenates, or distillation process used to make, for example, the gasoline they put in their cars. They just want their car to run well. While this would appear to make a refiner's problem easier, choosing their heterogeneous inputs, such as crude oil, satisfying federal, state and city environmental regulations, and all while maximizing profits, makes for an enormously complex optimization.

A.3 Estimation Algorithm

My estimation strategy involves matching utilization and investment moments. This requires that I solve for a policy function for each of these decisions and interpolate the functions to the realizations of the state variables in the data. The monthly utilization choice problem is a simple finite horizon dynamic program that I am able to solve by backward induction. So, for a given level of investment which induces a capacity for the plant, I can write the problem as:

$$\Pi_{iy} = Max_{\{u_{im}\}_{m=1}^{12}} \; E\left[\sum_{m=1}^{12} \mu^{m-1} \pi_{im}(u_{im}; x_{im}, \overline{q}_{iy})\right]. \tag{A.1}$$

Then, Π_{iy}, the aggregate discounted annual profit of the plant, becomes the payoff function for the infinite horizon problem. The Bellman equation for that problem

is:

$$V(x) \quad = \quad Max_r\Big\{\Pi_{iy}(r;x) + \delta V(x')P(x'|x,r)\Big\}. \tag{A.2}$$

To solve this equation, I could have used several different methods including successive approximations or collocation, but I chose policy function iteration, also known as the *Howard Policy Improvement Algorithm*. The first step is to guess a candidate policy function, which I call, $\sigma_t(x)$, where t indexes the iteration. Since this policy governs investment which affects optimal utilization, which in turn affects the probability of breakdown, I have to calculate the transition matrix given the policy: $P(x'|x,\sigma_t(x))$. Then comes the "policy evaluation step" which is to solve A.2, i.e.:

$$V_t(x) \quad = \quad [I - \delta P(x'|x,\sigma_t(x))]^{-1}\Pi_{iy}(\sigma_t(x);x). \tag{A.3}$$

For a size K state space, this involves the inversion of a KxK matrix which makes it difficult to estimate the with too fine of a discretization. With the value function in hand, I move to the "policy improvement step" which updates the policy function:

$$\sigma_{t+1}(x) \quad = \quad argmax_r \Big\{\Pi_{iy}(r;x) + \delta V_t(x')P(x'|x,r)\Big\}. \tag{A.4}$$

Finally, I compare $\sigma_{t+1}(x)$ to $\sigma_t(x)$ and repeat the process until convergence.

A.4 Additional Tables

Table A.2: Industry Concentration

	1970	1980	1991	2001	2004	2005	2006	2007	2008
US									
4-Firm (%)			31.4	40.2	44.4	43.0	45.8	44.1	41.2
8-Firm (%)			52.2	61.6	69.4	68.4	72.0	69.5	63.7
HHI			437.0	611.0	728.0	727.0	776.4	730.3	644.2
PADD 1									
4-Firm (%)			59.2	80.7	76.7	85.8	87.3	87.3	87.0
8-Firm (%)			88.7	99.0	97.9	99.4	99.4	99.4	99.4
HHI			1,225.0	2,158.0	1,943.0	2,505.0	2,537.5	2,540.2	2,524.7
PADD 2									
4-Firm (%)	38.3	37.4	39.3	50.9	57.1	57.1	59.6	55.5	50.5
8-Firm (%)	59.7	60.0	65.0	75.6	82.6	82.6	85.0	80.9	75.9
HHI			675.0	961.0	1,063.0	1,059.0	1,114.0	1,031.3	950.8
PADD 3									
4-Firm (%)	44.0	36.2	36.3	48.4	56.3	56.0	57.8	56.0	50.9
8-Firm (%)	64.8	54.5	58.5	66.5	78.8	78.2	81.2	77.6	73.2
HHI			578.0	851.0	1,018.0	1,005.0	1,052.2	976.7	909.2
PADD 4									
4-Firm (%)	53.5	48.0	55.8	58.1	46.1	45.7	50.9	50.7	58.7
8-Firm (%)	81.7	75.3	83.6	86.9	81.2	80.4	85.5	85.2	84.3
HHI			1,080.0	1,179.0	944.0	935.0	1,047.7	1,031.5	1,405.5
PADD 5									
4-Firm (%)	66.5	54.4	53.8	60.2	62.4	62.4	59.1	59.2	61.8
8-Firm (%)	95.2	76.5	74.2	86.9	92.7	92.8	89.5	89.6	89.4
HHI			965.0	1,148.0	1,246.0	1,247.0	1,162.2	1,168.7	1,195.7
California									
4-Firm (%)			58.9	68.7	66.2	66.5	62.3	62.5	63.0
8-Firm (%)			82.5	95.1	96.3	96.3	92.1	93.2	93.2
HHI			1,184.0	1,481.0	1,475.0	1,475.0	1,354.9	1,367.2	1,368.8
Gulf Coast									
4-Firm (%)							59.1	60.1	53.7
8-Firm (%)							83.5	83.1	76.7
HHI							1,107.9	1,110.5	995.0
PADDs 1 & 3									
4-Firm (%)	40.9	35.0	36.7	44.6	54.6	52.5	55.4	54.0	50.2
8-Firm (%)	62.3	55.0	57.2	65.3	76.1	75.5	79.5	76.6	72.8
HHI			561.0	741.0	919.0	890.0	967.9	991.1	861.2
PADDs 2 & 3									
4-Firm (%)			30.7	42.5	46.2	45.9	50.0	47.5	44.4
8-Firm (%)			56.5	64.9	75.6	75.2	79.9	76.2	70.3
HHI			455.0	681.0	826.0	818.0	894.6	822.7	742.9
PADDs 1, 2, & 3									
4-Firm (%)	35.2	30.7	30.2	39.4	45.9	44.5	49.2	47.1	43.9
8-Firm (%)	58.0	49.2	53.6	63.5	73.1	72.6	78.3	75.1	69.6
HHI			460.0	638.0	789.0	783.0	872.7	807.9	731.4

Source: EIA. Concentration based on operating capacity of crude oil distillation measured per calendar day on January 1st of the given year. The FTC generated the table through 2004 and I extended it through 2008. Upper Midwest: Illinois, Indiana, Kentucky, Michigan, and Ohio. Increase from 2004 to 2005 HHI's in PADDs I and III primarily due to the merger between Valero and Premcor. Capacities used in this table are at the corporate level (multiple refineries owned by the same corporation are aggregated).

Table A.3: Cost Estimates

Year	Parameter	Market 1 Coefficient	Market 1 Std. Err.	Market 2 Coefficient	Market 2 Std. Err.	Market 3 Coefficient	Market 3 Std. Err.
	Q (γ_0)	3.45***	0.01	0.36***	0.10	7.99***	0.75
	Q^2 (γ_1)	2.70***	0.01	10.86	11.18	5.45***	0.21
1995	$Q*P^c$ (γ_2)	0.29***	0.00	0.06***	0.02	0.28***	0.04
	Investment (γ_3)	4.41***	0.14	4.56	5.36	7.80	8.70
	$Investment^2$ (γ_4)	-4.41***	0.07	-2.99***	0.74	-5.52	5.01
	Q (γ_0)	3.48***	0.00	2.62***	0.38	0.05	2.09
	Q^2 (γ_1)	6.19***	0.01	5.21***	0.31	6.02***	0.44
1996	$Q*P^c$ (γ_2)	0.03***	0.00	0.03*	0.02	1.00***	0.03
	Investment (γ_3)	4.01***	0.15	5.58	51.23	3.84	11.82
	$Investment^2$ (γ_4)	-1.27***	0.05	-0.97	8.91	-2.09**	1.03
	Q (γ_0)	0.05*	0.03	0.92***	0.19	1.08	1.98
	Q^2 (γ_1)	5.14***	0.05	7.85***	0.15	7.30***	0.04
1997	$Q*P^c$ (γ_2)	0.08***	0.00	0.05***	0.01	0.38***	0.00
	Investment (γ_3)	4.25***	0.03	3.60**	1.64	8.88***	0.21
	$Investment^2$ (γ_4)	-0.81***	0.01	1.03	1.88	-1.86***	0.04
	Q (γ_0)	0.17***	0.03	0.05	26.36	1.16***	0.31
	Q^2 (γ_1)	1.00***	0.04	3.68	8.20	3.40***	0.24
1998	$Q*P^c$ (γ_2)	1.00***	0.01	0.02	55.93	0.86***	0.08
	Investment (γ_3)	-17.65	110.67	3.28	6.13	5.15	95.97
	$Investment^2$ (γ_4)	25.35	33.80	-4.30	32.06	-1.91	1.75
	Q (γ_0)	2.70***	0.04	0.44	51.57	6.94	35.07
	Q^2 (γ_1)	5.79***	0.18	2.13	6.43	7.35***	0.05
1999	$Q*P^c$ (γ_2)	0.01***	0.00	0.27	3.96	0.12	19.64
	Investment (γ_3)	4.65	14.90	5.90	11.03	9.53***	0.73
	$Investment^2$ (γ_4)	-0.82	1.31	-6.05	58.11	-0.92***	0.13
	Q (γ_0)	6.19***	0.57	0.04	0.19	10.29***	1.57
	Q^2 (γ_1)	5.89***	0.11	11.36***	0.63	6.36***	0.41
2000	$Q*P^c$ (γ_2)	0.00	0.00	0.00	0.00	0.01	0.06
	Investment (γ_3)	5.65*	4.16	4.08	4.40	11.85***	1.33
	$Investment^2$ (γ_4)	-2.82***	0.44	-0.99	2.13	5.26	9.43
	Q (γ_0)	0.32***	0.06	0.05***	0.01	0.03	2.92
	Q^2 (γ_1)	5.75***	0.06	23.84***	1.07	2.63**	1.19
2001	$Q*P^c$ (γ_2)	0.02***	0.00	0.00***	0.00	1.00***	0.20
	Investment (γ_3)	4.56***	0.53	3.91***	0.35	9.74	15.17
	$Investment^2$ (γ_4)	1.12***	0.07	-4.79***	0.36	-5.05***	0.99
	Q (γ_0)	2.24***	0.74	0.12***	0.03	0.58	0.52
	Q^2 (γ_1)	4.51***	0.10	3.70***	0.74	6.90***	0.58
2002	$Q*P^c$ (γ_2)	0.16***	0.03	0.98***	0.08	0.28***	0.05
	Investment (γ_3)	17.48**	9.18	5.49**	2.74	6.75	1,402.90
	$Investment^2$ (γ_4)	3.49	14.69	-1.09	0.86	-0.87	6.75
	Q (γ_0)	0.88***	0.18	13.42	394.71	0.03	0.22
	Q^2 (γ_1)	5.87***	0.11	0.56	27.99	4.50***	0.24
2003	$Q*P^c$ (γ_2)	0.08***	0.01	0.32	3.94	0.79***	0.04
	Investment (γ_3)	4.32***	0.70	5.43**	3.15	4.73***	1.64
	$Investment^2$ (γ_4)	2.75***	0.89	-1.02	1.88	-3.08*	2.14
	Q (γ_0)	3.18***	0.22	0.17***	0.07	0.15	0.45
	Q^2 (γ_1)	8.04***	0.13	28.65***	8.47	11.49***	0.68
2004	$Q*P^c$ (γ_2)	0.00***	0.00	0.01***	0.00	0.00	0.02
	Investment (γ_3)	7.48***	0.70	5.35***	1.19	7.07	7.23
	$Investment^2$ (γ_4)	2.09***	0.10	-5.14***	0.57	-2.84	4.96
	Q (γ_0)	0.34***	0.02	0.90***	0.11	0.04	34.02
	Q^2 (γ_1)	2.85***	0.05	8.52***	0.07	1.39	13.95
2005	$Q*P^c$ (γ_2)	1.00***	0.01	1.00***	0.01	1.00***	0.05
	Investment (γ_3)	10.42	24.93	11.69***	1.40	10.74***	4.42
	$Investment^2$ (γ_4)	2.05	1.60	-2.97***	0.71	-1.15	2.36
	Q (γ_0)	2.92***	0.06	0.01	0.19	1.01***	0.35
	Q^2 (γ_1)	1.39***	0.02	4.67***	0.93	4.79***	0.34
2006	$Q*P^c$ (γ_2)	1.00***	0.00	1.00***	0.03	1.00***	0.03
	Investment (γ_3)	9.44	443.89	8.42	7.81	7.43	493.22
	$Investment^2$ (γ_4)	2.85	134.06	0.01	130.15	0.15	157.43

****, **, * Significant at the 1%, 5%, and 10% level respectively.*

Appendix B

Chapter 2 Supplement

Top 20 Most Actively Searched Drugs

Drug Name	Num. Of Sessions	Num. Of Queries	Mean Queries Per Session	Ad Spending (Millions)
viagra	778	2,544	3.27	$80.56
lexapro	728	1,734	2.38	$1.18
depo	661	1,437	2.17	$0.00
xanax	583	1,497	2.57	$0.00
zoloft	566	1,305	2.31	$46.73
wellbutrin	489	1,193	2.44	$108.14
ambien	484	1,012	2.09	$130.20
cymbalta	477	1,060	2.22	$6.33
lyrica	430	886	2.06	$0.58
effexor	405	897	2.21	$4.05
insulin	384	1,127	2.93	$0.00
lipitor	384	754	1.96	$93.54
paxil	358	873	2.44	$0.11
prozac	330	757	2.29	$0.52
celebrex	290	744	2.57	$3.59
cialis	284	830	2.92	$110.94
seroquel	267	521	1.95	$2.16
lithium	265	767	2.89	$0.05
oxycontin	258	1,006	3.90	$0.00
toprol	253	493	1.95	$0.00
Total	**8,674**	**21,437**	**2.48**	**$588.66**

Ad spending is total expenditure on all forms of DTCA in 2005. These 20 drugs account for 30% of all search sessions, 33% of clicks, and 17% of DTCA spending.

Table B.1: 20 Most Actively Searched Drugs

Top 20 Most Actively Advertised Drugs

Drug Name	Num. Of Sessions	Num. Of Queries	Mean Queries Per Session	Ad Spending (Millions)
nexium	250	439	1.76	$226.34
lunesta	185	383	2.07	$215.14
vytorin	181	330	1.82	$155.26
crestor	226	441	1.95	$141.82
ambien	484	1,012	2.09	$130.20
nasonex	79	143	1.81	$124.16
flonase	65	113	1.74	$112.82
cialis	284	830	2.92	$110.94
lamisil	117	276	2.36	$110.51
plavix	199	371	1.86	$110.16
wellbutrin	489	1,193	2.44	$108.14
singulair	141	323	2.29	$105.05
lipitor	384	754	1.96	$93.54
imitrex	40	106	2.65	$82.21
viagra	778	2,544	3.27	$80.56
valtrex	161	307	1.91	$72.11
prevacid	154	242	1.57	$71.88
allegra	184	379	2.06	$71.04
boniva	87	178	2.05	$66.45
zelnorm	103	150	1.46	$62.45
Total	**4,591**	**10,514**	**2.10**	**$2,250.77**

Ad spending is total expenditure on all forms of DTCA in 2005. These 20 drugs account for 16% of all search sessions and clicks, and 65% of DTCA spending.

Table B.2: 20 Most Advertised Drugs

Variable Definition for OLS and Probit Models

Dependent Variables	Description
Total Sessions	total search sessions for a drug
Beyond Page 1	0/1 indicator; 1 if a user clicks on a link on page 2 or higher
Independent Variables	
age	years since FDA approval
dtca	total DTCA spending, available 1994 - February 2006, logs
alltv	total DTCA spending on TV, available 1994 - February 2006, logs
allmags	total DTCA spending in magazines, available 1994 - February 2006, logs
allnewsp	total DTCA spending in newpapers, available 1994 - February 2006, logs
allradio	total DTCA spending on radio, available 1994 - February 2006, logs
outdoor	total DTCA spending on outdoor media, available 1994 - February 2006, logs
internet	total DTCA spending on the internet, available 1994 - February 2006, logs
X_Y_qtrb4	total spending on X in the Y quarter prior to search, logs
rld	0/1 indicator; 1 if producer of drug is the innovator/pioneer

Note: Stock regressions include the total spending for a drug for all months between January 1994 and February 2006. Regressions involving previous quarter data include spending from December 2005 - February 2006. The depreciation analysis also includes spending from three previous quarters in 2005.

Table B.3: Variables Used in Regressions.

Appendix C

Chapter 3 Supplement

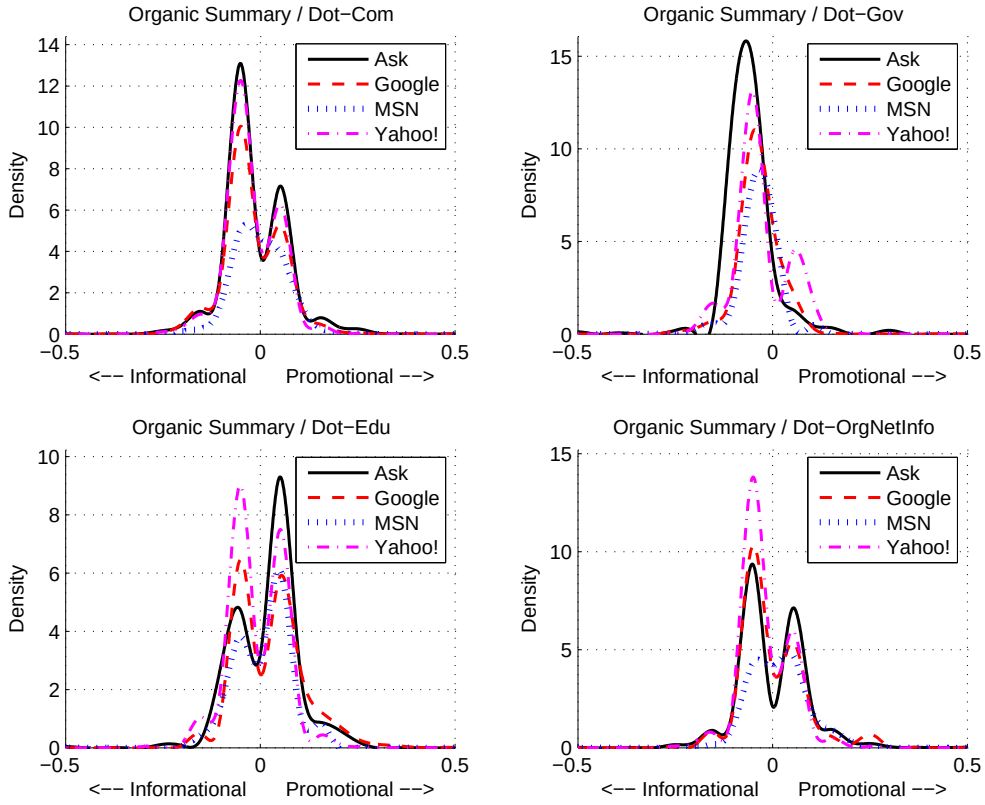


Figure C.1: Kernel Density of Summary Content, Organic Results, By Extension

Complete List of Search Queries

1 actos	59 actos price	117 crestor information	175 insulin	233 lipitor buy	291 metformin generic	349 pravachol cost	407 vivactil
2 actos "actos plus"	60 actos risks	118 crestor interactions	176 insulin "actos plus"	234 lipitor cheap	292 metformin glucophage	350 pravachol crestor	408 vytorin
3 actos "blood glucose"	61 actos sale	119 crestor lipitor	177 insulin "blood glucose"	235 lipitor cholesterol	293 metformin information	351 pravachol discount	409 welchol
4 actos "side effects"	62 actos side effects	120 crestor lovastatin	178 insulin "side effects"	236 lipitor cost	294 metformin insulin	352 pravachol generic	410 wellbutrin
5 actos "weight gain"	63 actos weight gain	121 crestor pravachol	179 insulin "weight gain"	237 lipitor crestor	295 metformin interactions	353 pravachol information	411 zetia
6 actos "weight loss"	64 actos weight loss	122 crestor price	180 insulin "weight loss"	238 lipitor discount	296 metformin price	354 pravachol interactions	412 zocor
7 actos actos plus	65 advicor	123 crestor risks	181 insulin actos	239 lipitor generic	297 metformin risks	355 pravachol lipitor	413 zocor "adverse effects"
8 actos blood glucose	66 altocor	124 crestor sale	182 insulin actos plus	240 lipitor information	298 metformin sale	356 pravachol lovastatin	414 zocor "blood pressure"
9 actos buy	67 amaryl	125 crestor side effects	183 insulin blood glucose	241 lipitor interactions	299 metformin side effects	357 pravachol price	415 zocor "side effects"
10 actos cheap	68 amitriptyline	126 crestor statins	184 insulin buy	242 lipitor lovastatin	300 metformin weight gain	358 pravachol risks	416 zocor adverse effects
11 actos cost	69 anafranil	127 crestor zocor	185 insulin cheap	243 lipitor pravachol	301 metformin weight loss	359 pravachol sale	417 zocor blood pressure
12 actos diabetes	70 avandamet	128 cymbalta	186 insulin cost	244 lipitor price	302 mevacor	360 pravachol side effects	418 zocor buy
13 actos discount	71 avandaryl	129 desyrel	187 insulin diabetes	245 lipitor risks	303 mirtazapine	361 pravachol statins	419 zocor cheap
14 actos generic	72 avandia	130 doxepin	188 insulin discount	246 lipitor sale	304 nardil	362 pravachol zocor	420 zocor cholesterol
15 actos glucophage	73 bupropion	131 duetact	189 insulin generic	247 lipitor side effects	305 niaspan	363 pravastatin	421 zocor cost
16 actos information	74 buspar	132 edronax	190 insulin glucophage	248 lipitor statins	306 norpramin	364 precose	422 zocor crestor
17 actos insulin	75 byetta	133 effexor	191 insulin information	249 lipitor zocor	307 nortriptyline	365 prozac	423 zocor discount
18 actos interactions	76 celexa	134 elavil	192 insulin interactions	250 lopid	308 novolin	366 prozac "long term"	424 zocor generic
19 actos metformin	77 celexa "long term"	135 endep	193 insulin metformin	251 lovastatin	309 novolog	367 prozac "side effects"	425 zocor information
20 actos plus	78 celexa "side effects"	136 escitalopram	194 insulin price	252 lovastatin "adverse effects"	310 nph insulin	368 prozac "weight gain"	426 zocor interactions
21 actos plus "blood glucose"	79 celexa "weight gain"	137 fenofibrate	195 insulin risks	253 lovastatin "blood pressure"	311 pamelor	369 prozac "weight loss"	427 zocor lipitor
22 actos plus "side effects"	80 celexa "weight loss"	138 fluoxetine	196 insulin sale	254 lovastatin "side effects"	312 parnate	370 prozac buy	428 zocor lovastatin
23 actos plus "weight gain"	81 celexa buy	139 fluvoxamine	197 insulin side effects	255 lovastatin adverse effects	313 paroxetine	371 prozac celexa	429 zocor pravachol
24 actos plus "weight loss"	82 celexa cheap	140 galvus	198 insulin weight gain	256 lovastatin blood pressure	314 paxil	372 prozac cheap	430 zocor price
25 actos plus actos	83 celexa cost	141 gemfibrozil	199 insulin weight loss	257 lovastatin buy	315 paxil "long term"	373 prozac cost	431 zocor risks
26 actos plus actos	84 celexa depression	142 glimepiride	200 januvia	258 lovastatin cheap	316 paxil "side effects"	374 prozac depression	432 zocor sale
27 actos plus blood glucose	85 celexa discount	143 glipizide	201 lantus	259 lovastatin cholesterol	317 paxil "weight gain"	375 prozac discount	433 zocor side effects
28 actos plus buy	86 celexa generic	144 glucophage	202 lescol	260 lovastatin cost	318 paxil "weight loss"	376 prozac generic	434 zocor statins
29 actos plus buy	87 celexa information	145 glucophage "actos plus"	203 lexapro	261 lovastatin crestor	319 paxil buy	377 prozac information	435 zoloft
30 actos plus cheap	88 celexa interactions	146 glucophage "blood glucose"	204 lexapro "long term"	262 lovastatin discount	320 paxil celexa	378 prozac interactions	436 zoloft "long term"
31 actos plus cheap	89 celexa lexapro	147 glucophage "side effects"	205 lexapro "side effects"	263 lovastatin generic	321 paxil cheap	379 prozac lexapro	437 zoloft "side effects"
32 actos plus cost	90 celexa long term	148 glucophage "weight gain"	206 lexapro "weight gain"	264 lovastatin information	322 paxil cost	380 prozac long term	438 zoloft "weight gain"
33 actos plus cost	91 celexa paxil	149 glucophage "weight loss"	207 lexapro "weight loss"	265 lovastatin interactions	323 paxil depression	381 prozac paxil	439 zoloft "weight loss"
34 actos plus diabetes	92 celexa price	150 glucophage actos	208 lexapro buy	266 lovastatin lipitor	324 paxil discount	382 prozac price	440 zoloft buy
35 actos plus diabetes	93 celexa prozac	151 glucophage actos plus	209 lexapro celexa	267 lovastatin pravachol	325 paxil generic	383 prozac risks	441 zoloft celexa
36 actos plus discount	94 celexa risks	152 glucophage blood glucose	210 lexapro cheap	268 lovastatin price	326 paxil information	384 prozac sale	442 zoloft cheap
37 actos plus discount	95 celexa sale	153 glucophage buy	211 lexapro cost	269 lovastatin risks	327 paxil interactions	385 prozac side effects	443 zoloft cost
38 actos plus generic	96 celexa side effects	154 glucophage cheap	212 lexapro depression	270 lovastatin sale	328 paxil lexapro	386 prozac weight gain	444 zoloft depression
39 actos plus generic	97 celexa weight gain	155 glucophage cost	213 lexapro discount	271 lovastatin side effects	329 paxil long term	387 prozac weight loss	445 zoloft discount
40 actos plus glucophage	98 celexa weight loss	156 glucophage diabetes	214 lexapro generic	272 lovastatin statins	330 paxil price	388 prozac zoloft	446 zoloft generic
41 actos plus glucophage	99 celexa zoloft	157 glucophage discount	215 lexapro information	273 lovastatin zocor	331 paxil prozac	389 questran	447 zoloft information
42 actos plus information	100 cholestyramine	158 glucophage generic	216 lexapro interactions	274 ludiomil	332 paxil risks	390 remeron	448 zoloft interactions
43 actos plus information	101 citalopram	159 glucophage information	217 lexapro long term	275 luvox	333 paxil sale	391 rezulin	449 zoloft lexapro
44 actos plus insulin	102 clozapine	160 glucophage insulin	218 lexapro paxil	276 metaglip	334 paxil side effects	392 sarafem	450 zoloft long term
45 actos plus insulin	103 colestid	161 glucophage interactions	219 lexapro price	277 metformin	335 paxil weight gain	393 sertraline	451 zoloft paxil
46 actos plus interactions	104 colestipol	162 glucophage metformin	220 lexapro prozac	278 metformin "actos plus"	336 paxil weight loss	394 serzone	452 zoloft price
47 actos plus interactions	105 crestor	163 glucophage price	221 lexapro risks	279 metformin "blood glucose"	337 paxil zoloft	395 simvastatin	453 zoloft prozac
48 actos plus metformin	106 crestor "adverse effects"	164 glucophage risks	222 lexapro sale	280 metformin "side effects"	338 pertofrane	396 sinequan	454 zoloft risks
49 actos plus metformin	107 crestor "blood pressure"	165 glucophage sale	223 lexapro side effects	281 metformin "weight gain"	339 prandin	397 starlix	455 zoloft sale
50 actos plus price	108 crestor "side effects"	166 glucophage side effects	224 lexapro weight gain	282 metformin "weight loss"	340 pravachol	398 strattera	456 zoloft side effects
51 actos plus price	109 crestor adverse effects	167 glucophage weight gain	225 lexapro weight loss	283 metformin actos	341 pravachol "adverse effects"	399 surmontil	457 zoloft weight gain
52 actos plus risks	110 crestor blood pressure	168 glucophage weight loss	226 lexapro zoloft	284 metformin actos plus	342 pravachol "blood pressure"	400 tofranil	458 zoloft weight loss
53 actos plus risks	111 crestor buy	169 glucotrol	227 lipitor	285 metformin blood glucose	343 pravachol "side effects"	401 tranylcypromine	
54 actos plus sale	112 crestor cheap	170 glucovance	228 lipitor "adverse effects"	286 metformin buy	344 pravachol adverse effects	402 trazodone	
55 actos plus sale	113 crestor cholesterol	171 glyburide	229 lipitor "blood pressure"	287 metformin cheap	345 pravachol blood pressure	403 tricor	
56 actos plus side effects	114 crestor cost	172 glyset	230 lipitor "side effects"	288 metformin cost	346 pravachol buy	404 trimipramine	
57 actos plus weight gain	115 crestor discount	173 humalog	231 lipitor adverse effects	289 metformin diabetes	347 pravachol cheap	405 venlafaxine	
58 actos plus weight loss	116 crestor generic	174 humulin	232 lipitor blood pressure	290 metformin discount	348 pravachol cholesterol	406 vestra	

Table C.1: List of Search Queries

Popular Keyword Lists

	Organic				Sponsored			
	Summary		Title		Summary		Title	
Obs	Promotional	Informational	Promotional	Informational	Promotional	Informational	Promotional	Informational
1	phentermine	effect	purchase	patient	canada	treatment	cheap	legal
2	pills	interactions	phentermine	oral	orders	natural	20mg	limited
3	shipping	including	shipping	lawsuit	cheap	tips	cost	promo
4	price	possible	save	statin	risk	options	sold	samples
5	viagra	serious	hosting	lawyer	hidden	out	500mg	cure
6	now	common	offer	hydrochloride	fees	anti	guaranteed	right
7	save	statin	cialis	withdrawal	fedex	depressant	canadapharmacy	injury
8	offers	oral	genuine	treatments	beat	expert	prescription	inhaled
9	lowest	cause	viagra	webmd	wholesale	breaking	better	avoid
10	cialis	patient	TRUE	lawyers	overnight	anxiety	30mg	possible
11	net	important	#3634	antidepressant	x30	birth	10mg	lawyer
12	fast	occur	#3619	heart	accredited	defects	pills	reviewed
13	cheapest	include	#3585	answers	ranked	use	safe	exubera
14	delivery	includes	delivery	handout	10mg	calcium	assistance	linked
15	tramadol	lowering	tramadol	information:	844	member	huge	damage
16	purchase	pdf	truly	safety	891	support	only	kidney
17	compare	provides	brand	rhabdomyolysis	brand	doctor	capsules	birth
18	easy	see	trusted	encyclopedia	onlineover	linked	100mg	take
19	blog	safety	sale	medlineplus	off	code	medicine	discovery
20	levitra	nausea	shop	class	x90	zip	rosuvastatin	attorney
21	pill	symptoms	home	niacin	discount	check	canadian	lower
22	soma	such	catalog	type	competitors	there	amp	dna
23	worldwide	warnings	fast	lowering	satisfaction	choose	savings	test
24	meds	risk	cost	suicide	customer	lawyer	canadadrugs	fatigue
25	link	learn	online:	attorney	1800	which	download	natural
26	pharmacies	muscle	xanax	lawsuits	pharmaciesfind	membersupport	comparison	hair
27	xanax	consumer	store	warnings	x60	proven	program	lamictal
28	homepages	over	easy	product	guaranteeaccredited	smarter	trusted	aid
29	snewman	following	top	precautions	500mg	questions	deals	performance
30	anti	know	guaranteed	antidepressants	beaten	really	iipitor	sexual
31	stmartin	experience	#3610	articles	fastdelivery	works	celexxa	lawsuit
32	discussionboard	levels	#1072	pcos	medisave	taking	45mg	drugs?
33	store	problems	compare	resistance	please	nutrition	meds	defect
34	aciphex	prescribed	pharmacies	learn	give	ebay	off	hypertension
35	great	help	link	syndrome	processing	join	pioglitazone	infant
36	prescriptions	usage	pump	koop	meds	know	direct	pulmonary
37	quality	hydrochloride	india	library	tablets	breakthrough	clearance	news
38	cost	prescribing	#3629	injury	fee	exciting	starting	defects
39	licensed	heart	topic	anxiety	available	one	today	contraceptive
40	day	precautions	usa	indications	x180	recommend	china	missed
41	products	potential	#3633	defective	convenient	most	directly	oral
42	valium	well	diabetic	mayoclinic	lowest	productsfind	incredibly	safe?
43	shop	sexual	aciphex	wikipedia	fda	solution	great	detailed
44	sale	diarrhea	name	blood	minutes	damage	rxdrugcard	prevachol
45	search	adverse	#3637	revolution	quality	loss	medication	resistance
46	make	clinical	bravenet	litigation	medication	statins	name	reverse
47	overnight	out	levitra	hydrobromide	sale	medicine	selling	locating
48	2006	many	#1086	ssri	ringtones	research	sale	need
49	posted	exercise	#3591	statins	30mg	contact	tickets	products
50	ultram	additional	xanga	oxalate	home	work	nordisk	review

Table C.2: Keywords Used in Classification Algorithm

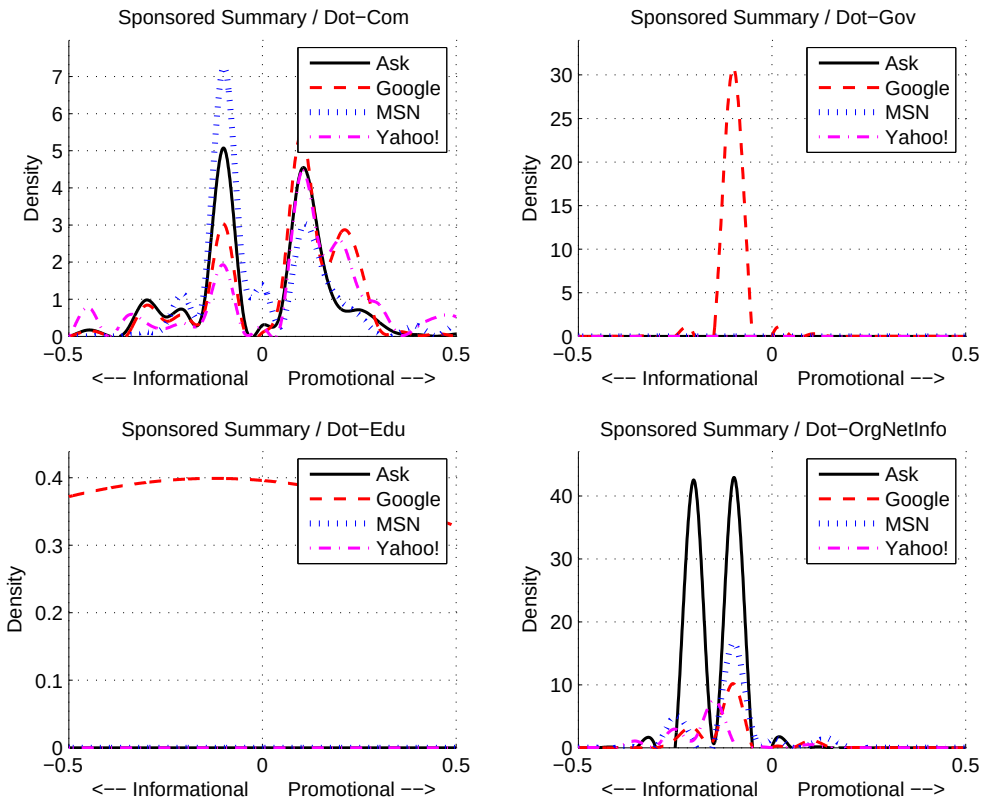


Figure C.2: Kernel Density of Summary Content, Sponsored Results, By Extension

Variable Summary for Probit Regression	
Dependent Variable	**Description**
Page1	0/1 indicator; 1 if the result appears on page 1 of the result
Independent Variables	
dtca_stock	total DTCA spending, 1994 - February 2007, billions
dotcom	0/1 indicator; 1 if dot-com
dotgov	0/1 indicator; 1 if dot-gov
dotedu	0/1 indicator; 1 if dot-edu
dotorgnetinfo	0/1 indicator; 1 if dot-org, net, or info
dotintl	0/1 indicator; 1 if dot-us, uk, or ca
dotcom_age	interaction term: dotcom and age
dotgov_age	interaction term: dotgov and age
dotedu_age	interaction term: dotedu and age
dotorg_age	interaction term: dotorg and age
dotcom_dtc	interaction term: dotcom and dtca_stock
dotgov_dtc	interaction term: dotgov and dtca_stock
dotedu_dtc	interaction term: dotedu and dtca_stock
dotorg_dtc	interaction term: dotorg and dtca_stock
prop_promo_summary	Proportion of words in the summary of organic links that are promotional
prop_info_summary	Proportion of words in the summary of organic links that are informational
prop_promo_title	Proportion of words in the title of organic links that are promotional
prop_info_title	Proportion of words in the title of organic links that are informational

Table C.3: Variable Definitions

References

[1] Aguirregabiria, V. P. Mira, (2006). "Sequential estimation of dynamic discrete games." *Econometrica,* 75(1), 2006.

[2] Athey, Susan and Glenn Ellison (working paper). "Position Auctions with Consumer Search."

[3] Attanasio, Orazio, (2000). "Consumer Durables and Inertial Behavior: Estimation and Aggregation of Ss Rules for Automobiles." *Review of Economic Studies,* October 2000.

[4] Bacon, Robert W., (1991). "Rockets and Feathers: The Asymmetric Speed of Adjustment of UK Retail Gasoline Prices to Cost Changes." *Energy Economics,* 13 July 1991.

[5] Bajari, Patrick, Lanier Benkard, and Jonathan Levin, (2007). "Estimating Dynamic Models of Imperfect Competition." *Econometrica,* 75(5), 2007.

[6] Battelle, John. (2005). The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture. Penguin Books Ltd, 2005.

[7] Benkard, Lanier, (2004). "A dynamic analysis of the market for wide-bodied commercial aircraft." *Review of Economic Studies,* 71(3), 2004.

[8] Besanko, David and Ulrich Doraszelski, (2004). "Capacity Dynamics and Endogenous Asymmetries in Firm Size." *The RAND Journal of Economics,* Vol. 35, No. 1. Spring 2004.

[9] Besanko, David A., Ulrich Doraszelski, Lauren Xiaoyuan Lu, and Mark A. Satterthwaite, (2008). "Lumpy Capacity Investment and Disinvestment Dynamics." Harvard Institute of Economic Research Discussion Paper No. 2154 Available at SSRN: http://ssrn.com/abstract=1117991.

[10] Borenstein, S., C. A. Cameron and R. Gilbert, (1997). "Do Gasoline Prices Respond Asymmetrically to Crude Oil Price Changes?" *Quarterly Journal of Economics,* 112(1), 1997.

[11] Borenstein, S., (1991). "Selling Costs and Switching Costs: Explaining Retail Gasoline Margins." *The RAND Journal of Economics,* 22(3), 1991.

[12] Borenstein, S., Andrea Shepard (1996). "Dynamic Pricing in Retail Gasoline Markets." *The RAND Journal of Economics,* 27(3), 1996.

[13] Chowdhury, A., G. Pass, C. Torgeson. (2006). "A Picture of Search" **The First International Conference on Scalable Information Systems**, Hong Kong, June, 2006.

[14] Day, Ruth. (2006). "Comprehension of Prescription Drug Information: Overview of A Research Program." *Proceedings of the American Association for Artificial Intelligence, Argumentation for Consumer Healthcare*. 2006.

[15] Day, Ruth. (2003). "Understanding Rx drug Information: TV ads, internet, hardcopy." *U.S. Food and Drug Administration*, Public Meeting on Effects of Direct-to-Consumer Advertising. 2003.

[16] Edelman, Ben and Michael Ostrovsky. (2007). "Strategic Bidder Behavior in Sponsored Search Auctions." *Decision Support Systems*, 2007.

[17] Energy Information Administration, US Department of Energy, (2007). "Refinery Outages: Description and Potential Impact on Petroleum Product Prices." March 2007.

[18] Energy Information Administration, US Department of Energy, (2008). "A Primer on Gasoline Prices." Online: `http://www.eia.doe.gov/bookshelf/brochures/gasolinepricesprimer/index.html` [Downloaded: 09/11/2008], May 2008.

[19] Ericson, Richard, and Ariel Pakes, (1995). "Markov-Perfect Industry Dynamics: A Framework for Empirical Work." *Review of Economic Studies,* 62:1, 53-83, 1995.

[20] Espey, Molly, (1996). "Explaining Variation in Elasticity of Gasoline Demand in the United States: A Meta Analysis." *The Energy Journal,* 17, 1996.

[21] The Federal Trade Commission, (2006). "Investigation of Gasoline Price Manipulation and Post-Katrina Gasoline Price Increases." Spring 2006.

[22] Ghose, A., and S. Yang. (2009). "An Empirical Analysis of Search Engine Advertising: Sponsored Search and Cross-Selling in Electronic Markets." Forthcoming in *Management Science*.

[23] Ghose, A., and S. Yang. (2008). "An Empirical Analysis of Sponsored Search Performance in Search Engine Advertising." **Proceedings of the ACM International Conference on Web Search and Data-mining Conference (WSDM 2008)**, Stanford, February 2008.

[24] Ghose, A., and S. Yang. (working paper). "Organic and Paid Search Advertising: Complements, Substitutes or Neither?"

[25] Goldberg, Pinelopi K. and Rebecca Hellerstein, (2008). "A Structural Approach to Explaining Incomplete Exchange-Rate Pass-Through and Pricing-to-Market." *The American Economic Review,* 98(2), 2008.

[26] The Government Accountability Office, (2006). "Energy Markets: Factors Contributing to Higher Gasoline Prices." GAO-06-412T. February 2006.

[27] Gron, Anne, Deborah Swenson, (2000). "Cost Pass-Through in the U.S. Automobile Market." *The Review of Economics and Statistics,* 82(2), 2000.

[28] Hamilton, James D., (1983). "Oil and the Macroeconomy since World War II." *The Journal of Political Economy,* 91(2), 1983.

[29] Hastings, Justine, Jennifer Brown, Erin Mansur, and Sofia Villas-Boas, (2008). "Reformulating Competition? Gasoline Content Regulation and Wholesale Gasoline Prices." *Journal of Environmental Economics and Management,* January 2008.

[30] Hotz, V. J., and R. A. Miller, (1993). "Conditional Choice Probabilities and the Estimation of Dynamic Models." *Review of Economic Studies,* 60:3, 497-529, 1993.

[31] Hubbard, Glenn, (1986). "Supply Shocks and Price Adjustment in the World Oil Market." *The Quarterly Journal of Economics,* 101(1), 1986.

[32] Huh, Jisu and Brenda Cude (2004). "Is the Information Fair and Balanced in Direct-to-Consumer Prescription Drug Websites?" *Journal of Health Communication,* 2004.

[33] ICF Consulting, (2005). "The Emerging Oil Refinery Capacity Crunch: A Global Clean Products Outlook." 2005.

[34] Jin, Ginger Zhe and Toshiaki Iizuka. (2005). "The Effects of Prescription Drug Advertising on Doctor Visits." *Journal of Economics & Management Strategy*, Fall 2005.

[35] Jin, Ginger Zhe and Toshiaki Iizuka. (2007). "Direct to Consumer Advertising and Prescription Choice." *Journal of Industrial Economics*, 2007.

[36] Knittel, Christopher, Jonathan E. Hughes, and Daniel Sperling, (2008). "Evidence of a Shift in the Short-Run Price Elasticity of Gasoline Demand." *The Energy Journal,* 29(1), January 2008.

[37] Kreps, David M., Jose A. Scheinkman, (1983). "Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes." *The Bell Journal of Economics,* 14(2), Autumn 1983.

[38] Lidderdale, T.C.M. (United States Energy Information Administration), (1999). "Environmental Regulations and Changes in Petroleum Refining Operations." Online: `http://www.eia.doe.gov/emeu/steo/pub/special/enviro.html` [Downloaded: 12/07/2007], 1999.

[39] Nielsen-online.com. (2009). "April 10, 2009 News Release." 2009.

[40] Nielsen-online.com. (2008). "The Second Opinion: How the Web Drives Healthcase Decisions." 2009. Webinar presented by Melissa Davies, September 3, 2008.

[41] Noel, Michael D., (2007). "Edgeworth Price Cycles, Cost-Based Pricing, and Sticky Pricing in Retail Gasoline Markets." *Review of Economics and Statistics,* Vol. 89, 2007.

[42] Pakes, Ariel, (2000). "A framework for applied dynamic analysis in I.O." Working paper no. 8024, NBER, Cambridge, 2000.

[43] Pakes, Ariel, Michael Ostrovsky, and Steven T. Berry, (2004). "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)." Harvard Institute. Economic Research Discussion Paper No. 2036, May 2004.

[44] Pakes, Ariel and P. McGuire, (1994). "Computing Markov-perfect Nash equilibria: Numerical implications of a dynamic differentiated product model." *Rand Journal of Economics,* 25(4), 1994.

[45] Peterson, D. J. and Sergej Mahnovski, (2003). "New Forces at Work in Refining: Industry Views of Critical Business and Operations Trends." Santa Monica, CA : RAND, 2003.

[46] Rust, John and Harry Paarsch, (forthcoming). "Valuing Programs with Deterministic and Stochastic Cycles." Forthcoming in the *Journal of Economic Dynamics and Control.*

[47] Rust, John, (2008). "Dynamic Programming." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008.

[48] Rust, John, (1987). "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher." *Econometrica,* 55:5, 999-1033, 1987.

[49] Ryan, Stephen, (forthcoming). "The Costs of Environmental Regulation in a Concentrated Industry." Forthcoming in *Econometrica*.

[50] Tirole, Jean, (1988). The Theory of Industrial Organization. Cambridge, MA: M.I.T. Press. 1988.

[51] The United States Senate, (2002). "Gas Prices: How are they Really Set?" Online: `http://www.senate.gov/~gov_affairs/042902gasreport.htm` [Downloaded 10/01/2007], May 2002.

[52] Varian, H. (2007). "Position Auctions." *International Journal of Industrial Organization,* 2007.