Senior Thesis

How much are secrets worth?
An asset pricing study on trade secret risks

Dazheng Percival Xu

May 2022

1 INTRODUCTION

Trade secrets are arguably a firm's most valuable intangible assets. They include secret recipes, proprietary technology, customer information, product designs, and business plans, etc. Firms use trade secrets to differentiate their products from competitors and make corporate decisions. According to Almeling (2012), trade secrets make up 80% of the value of S&P 500 companies.

While the importance of trade secret is clear, whether firms can keep their secrets is a completely different matter. The intangible nature of trade secrets means that the most treasured information can probably also be transferred with the simplest flash drive, not to mention employees that carry trade secrets with them when they move to competing firms. Trade secrets thefts are committed from both within the US and overseas, and they are estimated to cost US companies approximately 1 to 3 percent of the economy's GDP. (Ettredge et al. (2017)).

The past decade has seen a dramatic rise in trade secret thefts, and such thefts often inflict significant costs to company shareholders. On August 16, 2021, the US telecommunications giant T-Mobile (Ticker: $TMUS) suffered a severe data breach. Millions of customer names, SSNs, addresses, DOBs, and ID information were compromised. By the time market closed on that day, T-Mobile share price had tumbled over 3% and losses extended weeks after the incident (Figure 1.1).

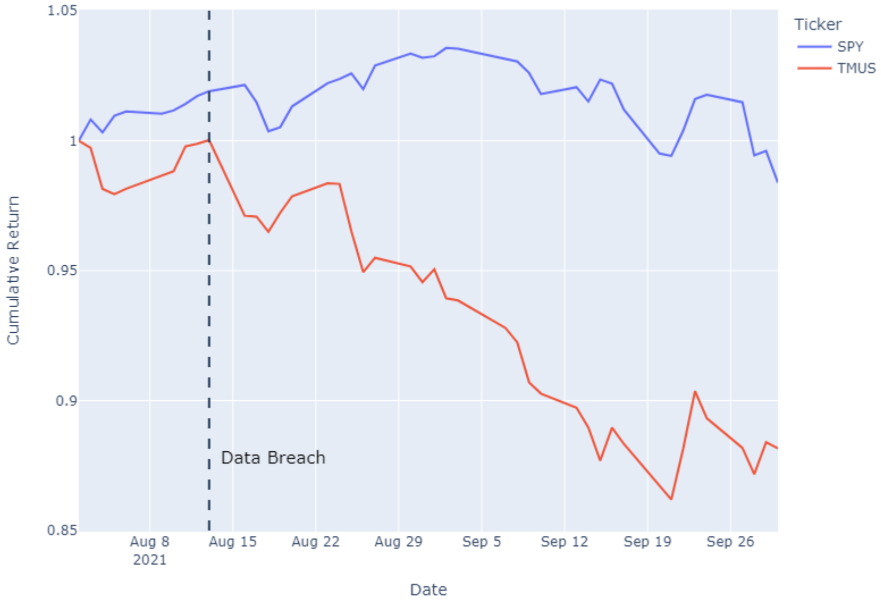

Figure 1.1: T-Mobile stock performance following data breach

Anecdotal evidence suggests that a firm's ability to protect trade secrets likely affect its future performance - but in which direction and what capacity? The capital asset pricing theory suggests that firms with high exposure to trade secret would carry a "trade secret risk", and investors shall be compensated with a premium for bearing this risk. This study tests this theory and seeks to add a potential risk factor to the asset pricing literature. This is important because it offers new insights to stock price drivers and has implications for market efficiency.

While it is easy to see the importance of trade secret risks, measuring this risk is far from easy. Trade secrets are by definition highly classified information, and companies do everything in their power to make sure they stay hidden. This poses significant challenges for trade secret measurements. Although obtaining precise firm-level data on trade secrets is an impossible task, it is possible to approximate it. Public companies in the US are obliged by law to disclose financial statements to investors, namely 10-K and 10-Q reports. Those reports contain information on the company's evaluation on its outlooks and risk factors, including the trade secret risk. By capturing the frequencies of certain keywords that appear in financial statements, we can produce a reasonable proxy for a company's exposure to trade secret risks.

In this study, I propose trade secrets exposure as a new pricing factor to explain the cross-sectional returns of the US stock market. By collecting 10-K reports and conducting textual analyses with natural language processing, I test the ability of a firm's exposure to trade secret risks as a factor to predict stock market returns. This is measured using two methods, namely portfolio sorting and the Fama-MacBeth regression.

I find a firm's exposure to trade secrets measured by keyword frequencies in its annual reports to be significant in explaining its stock returns in Fama-MacBeth regressions. On the other hand, portfolio sorting shows trade secret alone is not effective in generating excess returns.

The paper is structured as follows. The next section provides an overview of past literature on trade secrets and asset pricing. Section 3 reports the summary statistics and empirical strategy used in my study. Section 4 presents the main findings, and section 5 concludes the paper with final remarks.

2 Literature Review

Asset pricing has been and continues to be one of the most studied topics in empirical finance. At the core, asset pricing theories attempt to identify factors that contributes to determining an asset's price. Asset pricing theories have important implications because arbitrageurs could utilize them to profit from mispriced assets and improve market efficiency.

Since the seminal capital asset pricing model from Sharpe (1964), the field of asset pricing has discovered numerous pricing factors. The foundational capital asset pricing model from Sharpe (1964) traces an asset's return to two sources: systematic risks and idiosyncratic risks.

Later scholars further inspected an asset's idiosyncratic risks and have proposed various factors. A brief list of traditional factors includes low-volatility (Black (1972)), price-to-dividends ratio and price-to-earnings ratio (Campbell and Shiller (1988); Campbell and Shiller (1998)), momentum (Jegadeesh and Titman (1993)), quality (Sloan (1996); Asness et al. (2013)), and perhaps most famously the three-factor model (Fama and French (1992)) that combines the one-factor model from Sharpe (1964) with size and book-to-market ratio. Some alternative novel factors have been identified more recently, such as geographical dispersion (Garcia and Norli (2012)), innovative efficiency (Hirshleifer et al. (2012)) and cyber risks (Jamilov et al. (2021)).

While these alternative factors draw from an impressively wide range of fields, trade secret risks as a potential pricing factor has been overlooked by past literature. Prior to 2010, trade secret had received more academic attention in law than economics. Two laws stand out as centerpieces of trade secret laws in recent years: The Uniform Trade Secret Act (UTSA) and the Inevitable Disclosure Doctrine (IDD). Both the UTSA and the IDD are designed to protect a firm's interests when dealing with trade secrets. The UTSA was adopted by various states starting from 1981 and seeks to universalize trade secret laws across states, which historically differed at the state level. This ensures companies that operate in more than one state enjoy more trade secret protection as potential thefts cannot target one single state as a "weak spot" where trade secret laws are less present. Png (2017) found this law to be associated with more R&D among US companies, and this effect is especially pronounced in large companies and firms in the high-technology sector.

While the UTSA focuses on generalization, the Inevitable Disclosure Doctrine grants firms a specific avenue to protect their trade secrets. The IDD allows employers to enjoin former employees in procession of trade secrets from joining a competing firm because the former employee will "inevitably" leak trade secrets to her new employer. Companies frequently invoke IDD to protect their trade secrets and reactions from the courts have been mixed. Consequently, the effects of IDD are hard to grasp. Some researches have shown the IDD to increase firms reporting transparency, attract more venture capitals, and promote more aggressive capital structures by encouraging companies to take higher leverages. (Callen et al. (2020); Castellaneta et al. (2016); Klasa et al. (2018))

Although trade secrets as a standalone concept has not been used as a pricing factor, a closely related concept - patents - have received significant interest from asset pricing studies. Hsu (2009) used aggregate patent and R&D data as a proxy for technological innovation in an economy and found them to be predictive of future returns on the S&P 500 index. This result is not only applicable to the US market but supported by evidence from the international market as well. While Hsu (2009) focused on the aggregate level, Hirshleifer et al. (2012) broke down to the idiosyncratic firm level and found patents granted and patent citations data by individual companies can explain future returns. It is important, however, to note the difference between trade secrets and patent rights. As Png (2017) pointed out, patents provide broad exclusivity but have an expiration date and requires public disclosure. On the other hand, trade secrets can theoretically be kept forever and includes items such as customer lists and business plans that cannot be patented. By viewing trade secrets as a separate concept, I attempt to capture elements beyond patent rights that might also explain an asset's returns.

Finally, textual analysis has become a commonly used technique in academic research with the rapid development of natural language processing. Company filings have become exponentially lengthy after 2010 and quite frequently filled with incomprehensible corporate language. While it is impossible for a human to read and process information effectively from those documents, unprecedented computational power has given scholars a new option. Academics have used text-based analyses on earnings calls Jamilov et al. (2021), annual filings (Cohen et al. (2020)), quarterly filings (Ben-Rephael et al. (2021)) and more to extract valuable datasets that are otherwise unobtainable. In this research, I use company annual filings, commonly known as the 10-K, to conduct textual bigram searches. The following section discusses this process in more detail.

3 Empirical Tests

3.1 Data

I use two datasets in this research: company annual reports and stock price data at the individual firm level. Company annual reports, commonly known as the 10-K, are filed by publicly traded firms at their fiscal year end by regulation under the U.S. Securities and Exchange Commission (SEC). The 10-K is an ideal source of trade secret information because companies disclose their business model and future outlook in this report. If trade secret is important to a company, one can expect this firm to mention trade secret and measures of protecting them frequently in its 10-K.

The SEC manages EDGAR, a digital archive that stores all company filings including the 10-Ks. Starting on May 6, 1996, all public companies are required by law to file electronically to EDGAR. This makes EDGAR an ideal source for complete and well-documented company filings. I used a automated script to download and store the 10-K, 10-K/A, 10-K405, 10-KSB, 10-KT, 10KSB, 10KSB40, and 10KT405 reports from each firm from 1994 to 2016, totaling 231,111 files.

For stock market data, I use the CRSP/Compustat database to retrieve daily stock prices for every company listed on AMEX, New York Stock Exchange and Nasdaq. I include all companies listed at a given time instead of simply tracing back stock prices of companies listed as of now to prevent survivorship bias. In total, there is daily stock price data for XXX companies from 1994 to 2016.

3.2 Measuring trade secret risks

I estimate a company's trade secret risk in a given year by following a textual analysis procedure similar to Klasa et al. (2018). First, I look for the 10-K report filed by that company in the desired year. This 10-K report then serves as the corpus for the analyses that follow. The corpus is "tokenized" by transforming every word in the corpus into a "token". Tokens that are stop words (common words that do not carry contextual meaning such as *a*, *the*, *is*, etc.) are deleted to save computational time and space.

The remaining corpus consisting of tokens that are not stop words then goes through word by word inspection in search of matching keywords. This is a two-step process, each with a list of keywords. The first list is related with trade secrets, and the second list is related with protection. The word lists are directly replicated from Klasa et al. (2018). Although patents, trademark, and copyright are not directly related to trade secrets, they are included in this list because companies that are concerned with those concepts are also likely to be concerned with trade secrets.

List 1: Secret Words	List 2: Protect Words
intangible	
patent	
trademark	protect
copyright	protection
trade secret	safeguard
confidential information	lawsuit
proprietary information	
intellectual property	

Table 3.1: The two lists containing keywords searched in a corpus

Each company filing start with 0 keyword mentions. When traversing through tokens, I compare the current token with each keyword in the trade secret keyword list. If there is a match, I compare every neighboring token within a 20 tokens range with each keyword in the protection keyword list. If there is a match, 1 is added to the keyword mention count. This two-step process is then repeated for the next token until every token has been inspected. A firm's exposure to trade secret risks defined by the frequency of keyword appearances in a

20-word neighboring window around certain phrases, or a firm's "trade secret risk" in a given year is then calculated by:

$$\text{Trade Secret Risk} = \frac{\text{Total Mentions}}{\text{Total Filing Length (Number of Tokens)}}$$

Such a measure is based on the assumption that if a company is vulnerable or perceives itself as vulnerable against trade secret infringements, it would mention means to protect its secrets in their 10-K reports. If a company mentions protecting trade secrets frequently, then it is likely this company is subject to a higher degree of trade secret risks. Admittedly, this "neighboring word window" measurement is incapable of capturing complicated sentence structures in the English language. Yet, on average we should expect it to be a reasonable proxy for a company's trade secret risk.

3.3 10-K and 10-Q Reports

One common practice in natural language processing research in finance is to interpret not only 10-K filings but 10-Q filings as well. Similar to 10-K, 10-Q is a report filed by public companies that disclose information on the current standing of a firm. One difference between a 10-K and a 10-Q is the filing frequency: a company files 10-K reports annually and 10-Q reports quarterly. Another perhaps more important difference between them is a 10-K report is notably more detailed and include more sections than a 10-Q report. As a result, trade secret keywords are far more likely to appear in a 10-K than in a 10-Q. Although including 10-Q reports increases the sampling frequency, it might create unwanted bias in the dataset for the lack of discussion on trade secrets. For this reason I exclude 10-Q reports from my study and focus solely on 10-K reports.

3.4 Portfolio Formation

During the backtesting period, quarterly rebalanced portfolios are formed based on trade secret risks. The stock selection universe is all common stocks listed in the US stock market on AMEX, NYSE, and NASDAQ. At the beginning of each quarter (on the first business day of January, April, July, and October), stocks are ranked based on their trade secret risk factor and sorted into n portfolios. Stocks carrying the highest trade secret risk are firms that had the highest number of keyword mentions in the trailing 12 months. For example, the rebalance in January 2010 considers all firms that had filed a 10-K from January 2009 to January 2010 and rank them based on the trade secret risk calculated with their filings. On April 2010, stocks of firms that last filed a 10-K in the January 2009 to April 2009 period become ineligible and is replaced by stocks of firms that last filed a 10-K in the January 2010 to April 2010 period. However, since firms do not typically change their financial calendar the stock selection universe should by and large remain the same in practice. This rebalancing schedule is designed to always incorporate the latest information disclosed in company 10-Ks. If $n = 5$, then those stocks are grouped into 5 portfolios on the 20th, 40th, 60th, 80th, and 100th percentiles. Stocks within the same portfolio are equally weighted. The performance of those portfolios represent the hypothetical return of a buy-and-hold strategy that buys on a rebalance date and sells on the next rebalance date.

The top portfolio carries the highest trade secret risk, and the Capital Asset Pricing Model suggests it should provide a premium to investors to compensate for this risk. Conversely, the bottom portfolio bears the lowest trade secret risk and thus has the lowest trade secret risk premium. On each rebalance date, a zero-cost portfolio that takes a long position in the top portfolio and a short position on the bottom portfolio is constructed. The return of this portfolio, r_p, is equivalent to the difference of returns of the top and bottom portfolios and therefore an appropriate measure for the trade secret risk premium. If r_p is significantly different from zero, then there is some evidence that the top portfolio indeed carries a trade secret risk premium.

The backtesting period runs from July 1997 to December 2016. This starting date is chosen because filing to SEC EDGAR electronically became mandatory starting on May 1996, and July 1997 is the earliest rebalance date that allows a full year of look back period which contains all publicly traded companies.

3.5 Fama-MacBeth Regression

While portfolio sorting is commonly used in empirical finance to test a pricing factor's alpha-generating power, it does not show how much of the alpha originates from the factor of interest. This is a minor issue in empirical finance, but an unsatisfactory one to those who would like to precisely attribute asset returns to each individual factors. For this purpose, we have the Fama-MacBeth regression. First developed in Fama and MacBeth (1973), the Fama-MacBeth regression is a two-step regression. In the first step, a time-series regression is ran on the returns of each asset i against the proposed risk factors $f_1, f_2, ... f_n$. The estimated coefficients ($\hat{\beta}$s) on the factors, also known as "factor loadings", are then used as independent variables in the second step.

$$r_i = \alpha_{i,t} + \beta_{1,t} * f_{1,t} + ... + \beta_{n,t} * f_{n,t} + \epsilon_{i,t}$$

In the second step, a cross-sectional regression is ran for each time period t against the returns of each asset $i$in period t. The mean of estimated coefficients on the factor loadings for each period t, or $\mathbb{E}[\hat{\lambda}]$, are the "risk premia" associated with the risk factors in the model.

$$r_i = \lambda_{1,t} * \hat{\beta}_1 + ... + \lambda_{n,t} * \hat{\beta}_n + u_{i,t}$$

In this paper, I include five pricing factors in the Fama-MacBeth regression. The first three factors, market, size, and value are directly taken from Fama and French (1992). The fourth factor is momentum, and the fifth factor is my factor of interest - the "trade secret risk". The estimated premia are included in Section 4.

Table 3.2 reports the number of companies that filed at least once in a given year and the average number of words in a filing. It is worth noticing that the average number of words in a firm's filing has been on a steady RISE. The report length increased by 387% in the span of 20 years, and much of this increase in made of corporate language that are intentionally confusing and ambiguous. It is virtually impossible for a human to read those reports in large amounts, making automated textual analysis the only viable option.

Year	Number of Companies	Average Words
1995	3166	8299.4
1996	5316	9673.3
1997	5091	10458.4
1998	4264	11418.1
1999	2074	10005.1
2000	3815	11790.2
2001	5092	13780.1
2002	4804	16130.6
2003	4512	17346.7
2004	3863	19729.4
2005	3942	21819.4
2006	3597	24566.9
2007	3932	24206.3
2008	3974	25348.6
2009	3798	26271.4
2010	3668	26899.5
2011	3584	27634.1
2012	3533	28710.2
2013	3569	30903.7
2014	3642	31135.8
2015	3556	32129.7

Table 3.2

Table 3.3 describes trade secret keyword ratios and the proportion of firms with keyword mentions. The columns are defined as:

Year	Secret Ratio ‰	Protection Ratio ‰	Secret Firms %	Protection Firms %
1995	0.0027	0.0000	65.70	24.10
1996	0.0038	0.0000	71.67	31.04
1997	0.0040	0.0000	73.74	33.71
1998	0.0040	0.0000	74.34	34.10
1999	0.0043	0.0000	70.49	35.25
2000	0.0058	0.0000	78.03	39.53
2001	0.0107	0.0000	89.87	41.75
2002	0.0107	0.0000	88.82	42.36
2003	0.0099	0.0000	89.32	43.20
2004	0.0088	0.0000	88.79	44.03
2005	0.0094	0.0000	91.02	47.49
2006	0.0094	0.0000	91.49	48.71
2007	0.0105	0.0003	92.57	50.89
2008	0.0125	0.0004	94.64	52.62
2009	0.0116	0.0005	95.00	53.13
2010	0.0119	0.0006	96.16	54.14
2011	0.0120	0.0007	96.12	56.00
2012	0.0130	0.0008	96.55	57.26
2013	0.0127	0.0009	96.55	58.84
2014	0.0133	0.0011	97.09	61.39
2015	0.0140	0.0013	97.69	63.86

Table 3.3

- **Secret Ratio ‰**: The median of the ratio of secret keywords and total length of a firm's 10-K report.
- **Protection Ratio ‰**: The median of the ratio of protection keywords and total length of a firm's 10-K report.
- **Secret Firms %**: The percentage of firms with at least one secret keyword mention in a financial year, among all firms.
- **Protection Firms %**: The percentage of firms with at least one protection keyword mention in a financial year, among all firms.

Table 3.3 shows the rise of trade secrets in two dimensions. The first dimension - columns **Secret Firms%** and **Protection Firms %** - shows a steady increase in the proportions of firms with keyword mentions. Starting with 66% of the firms with at least one secret keyword mention and 24% with at least one protection keyword mention in 1995, the percentages have increased to 98% and 64% by the end of 2015. It is clear that on an overall level, more firms are viewing protecting their trade secrets as an important matter.

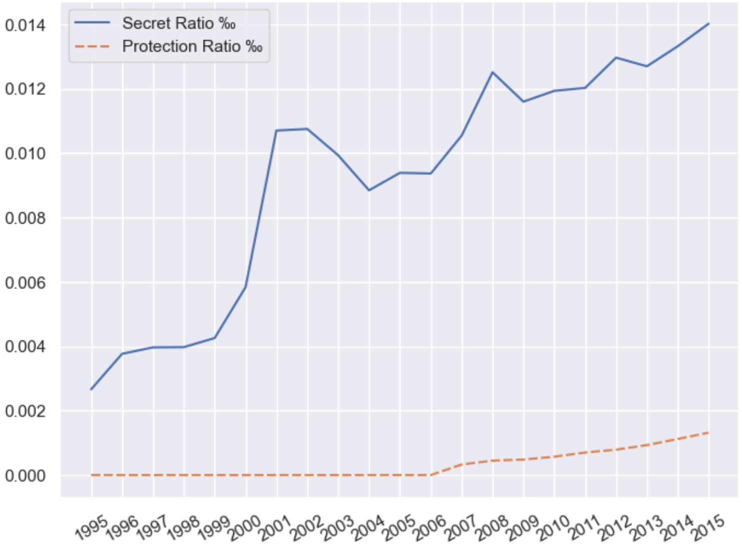

Figure 3.1: Median Secret and Protection Ratios by Financial Year (‰)

Columns **Secret Ratio ‰** and **Protection Ratio ‰** shows the second and more implicit dimension of the growth in trade secret risks. As illustrated in Figure 3.1, the increase in the ratio of keyword mentions to total lengths are not at all smooth. It is important to distinguish that while the first dimension asks *if* firms care about trade secrets, the second dimension asks *how much* firms care about them. This is especially obvious in the column **Protection Ratio ‰**. From 1995 to 2006, the median of this ratio remained at zero. In other words, at least half of the firms did not consider trade secret risk as a substantial threat until 2006. After that, a clear linear trend in the protection ratio can be seen. This point is better shown in Figure 3.2, as frequency distributions up to 2006 are heavily right-skewed.

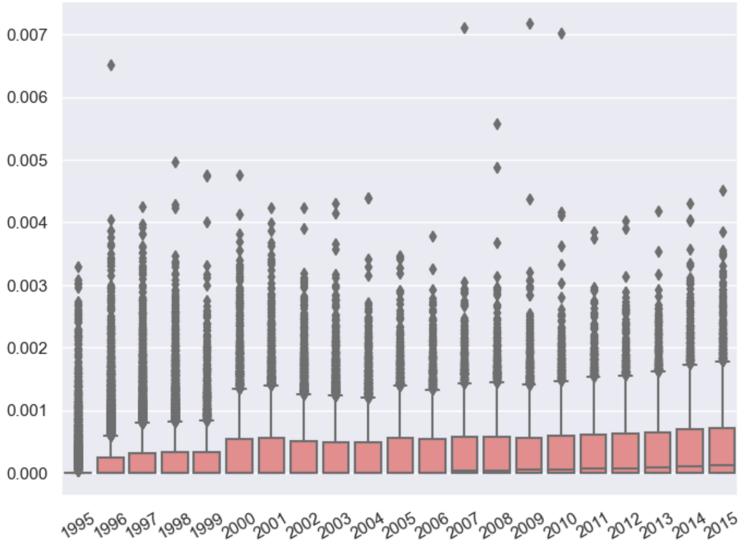

Figure 3.2: Distribution of protection ratios by financial year

Figure 3.2 shows another interesting story. While the distributions are heavily skewed, they also have extremely long tails. Statistical outliers have ratios that are several times higher than their peers. In other words, some companies mention protection keywords significantly more than others. Naturally, the next question is: Who are those firms? One reasonable hypothesis would be those firms belong to one or more groups which are more sensitive to trade secrets than others. An intuitive way to group firms is to categorize them into different industries. I grouped each firm into one of the nine sectors according to the Standard Industry Classification defined by the SEC (excluding the "Nonclassifiable" category). In addition, each firm is also associated with an industry - a subcategory of a sector. For example, Apple Inc. belongs to the electronic computers industry in the manufacturing sector.

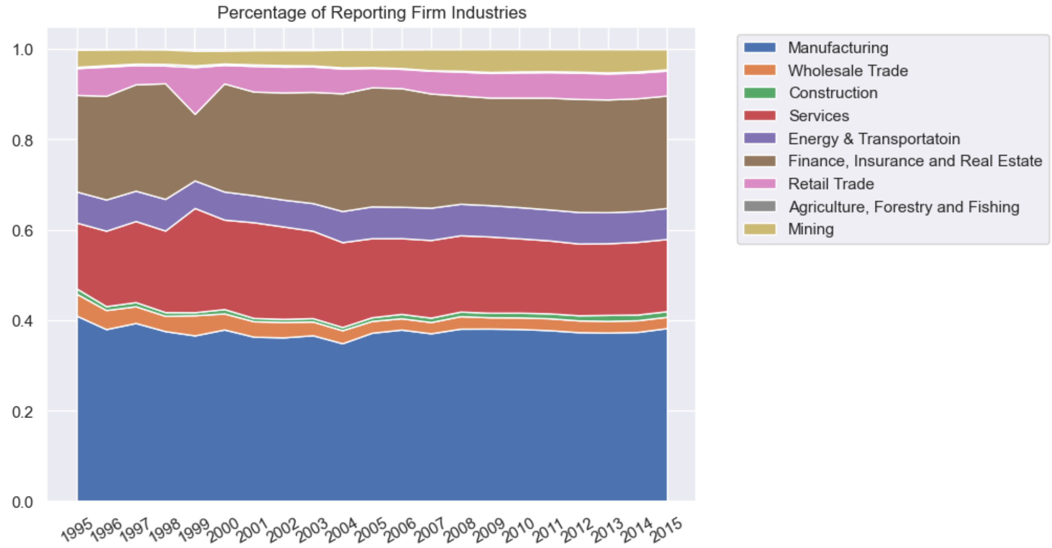

Figure 3.3: Percentage of firms in each industry by financial year

Figure 3.3 illustrates the percentage of each sector in the sample period. The size and rank of each sector remained largely unchanged: the Manufacturing sector contains approximately 40% of the firms in the sample, followed by the Financial sector with about 20% and the Services sector with about 15%. Figure 3.4 and Figure 3.5 present the average number of keywords within each sector over the sample period. Manufacturing is the sector with the most keyword mentions, followed by Services. Two particularly interesting sectors are Financials and Agriculture. The former is the third-largest sector in the sample but does not mention trade secrets often in its annual reports. Conversely, Agriculture make up less than 1% of all firms in the sample but mention trade secrets more than all sectors but Manufacturing and Services.

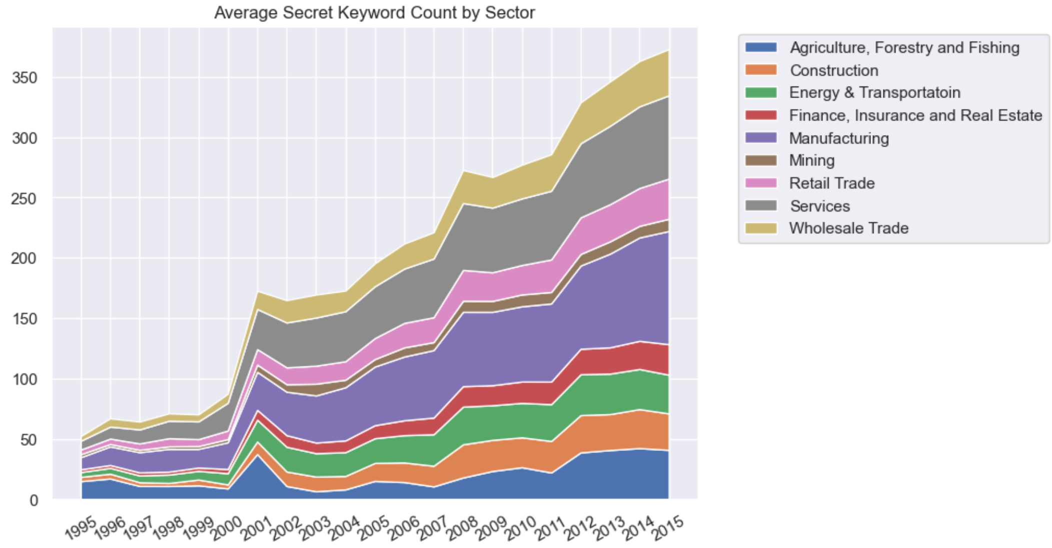

Figure 3.4: Average number of secret keywords by sector

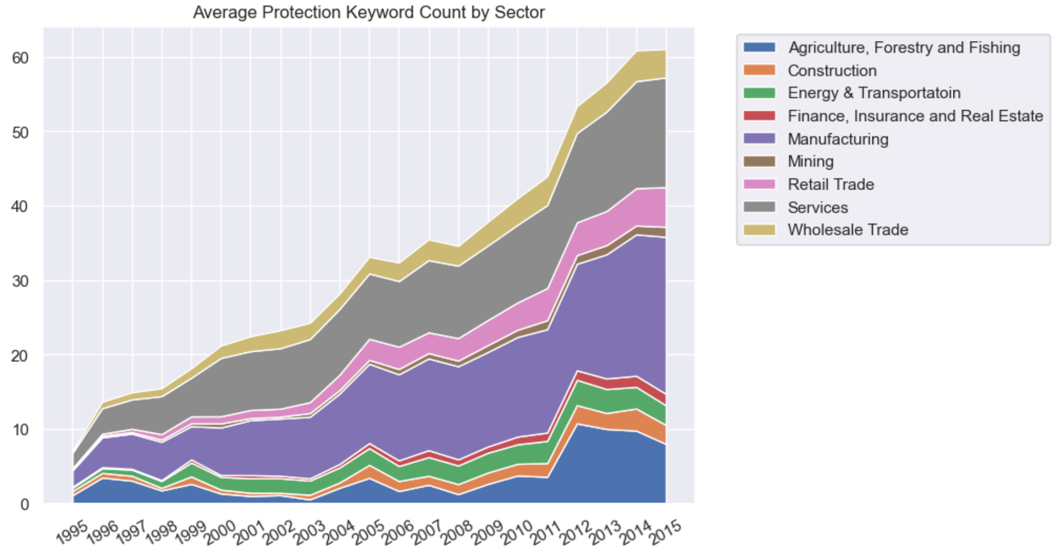

Figure 3.5: Average number of secret keywords by sector

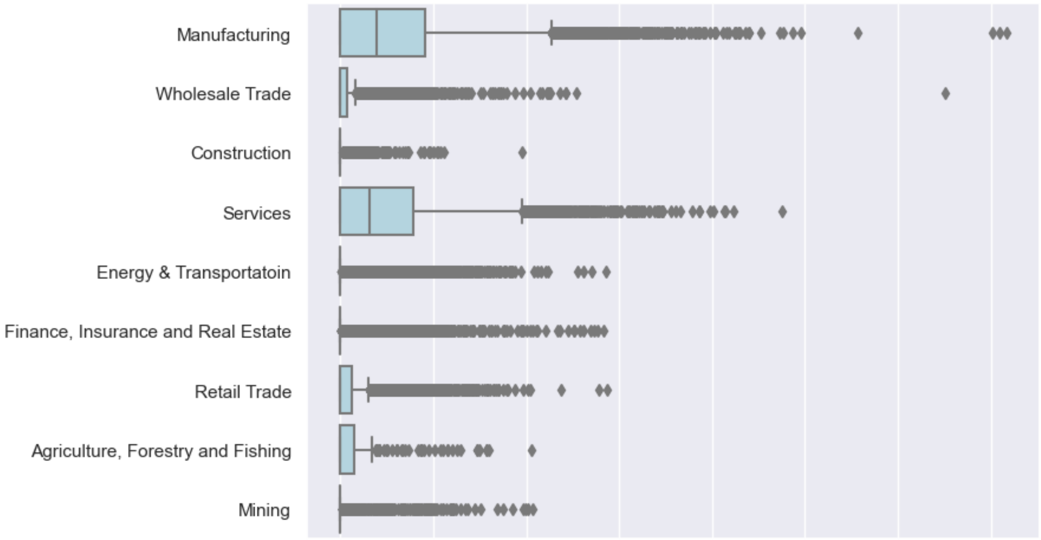

Figure 3.6: Distribution of protection ratios by sector

Finally, Figure 3.6 shows the distribution of protection keyword ratios by industry. It is immediately obvious that Manufacturing and Services have substantially less right-skewed distributions compared to the rest of the sectors. At the same time, the outliers in those two sectors also exhibit higher values than outliers in other sectors. This is further proved by Table 3.4 which reports the top ten industries with the most number of protection keyword mentions over the sample period. All ten industries are related to software services or medical/pharmaceutical items. This is perhaps unsurprising as trade secrets such as recipes and customer lists are vital to software and pharmaceutical companies because they are closely connected to the product offered. This gives us a clearer picture of who the outliers are in Figure 3.2. Trade secret is highly sensitive to the manufacturing and services sectors, and firms in those sectors are most likely to be the main driver of the overall increase of keyword mentions among all firms.

Industry	Sector	Secret Count	Protection Count
Pharmaceutical Preparations	Manufacturing	359379	86483
Prepackaged Software	Services	152344	41203
Semiconductors & Related Devices	Manufacturing	121225	31479
Biological Products	Manufacturing	105456	24338
Surgical & Medical Instruments & Apparatus	Manufacturing	82425	20555
Business Services	Services	86597	20350
Commercial Physical & Biological Research	Services	39721	10479
Computer Integrated Systems Design	Services	46158	10316
Electromedical & Electrotherapeutic Apparatus	Manufacturing	40826	10033
In Vitro & In Vivo Diagnostic Substances	Manufacturing	36278	9057

Table 3.4: Total number of keywords by industry

4 Main Findings

4.1 Portfolio Sorting

Portfolio sorting indicates trade secret risk alone does not generate meaningful excess returns. I test portfolio sorting on 3 samples: All firms, firms in the Manufacturing industry and firms in the Services industry. I use the secret ratio (SRT) and protection ratio (PRT) to sort firms into sub-portfolios, and construct long-short portfolios with long positions in the top portfolio and short positions in the bottom portfolio. Moreover, I also use raw factors divided by their industry medians to account for differences in industry-wide trade secret risks. The performance of the hypothetical portfolios are presented in Table 4.1.

Based on those results, one can conclude trade secret risk by itself is not effective in generating excess returns. The returns for all hypothetical portfolios are virtually zero with low significance levels.

4.2 Fama-MacBeth Regression

The Fama-MacBeth regression tells another story. I run three Fama-MacBeth regressions with different factors: Regression (1) treats the trade secret risk factor (PRT) as the only fac-

Factor	Firms	Annualized Return	Annualized Volatility	Risk-Adjusted Return	Significance (t)
SRT	All	0.0072	0.0937	0.0768	0.6608
PRT	All	-0.0064	0.0938	-0.0682	-0.1177
SRT / Med(I)	All	0.0004	0.0653	0.0061	0.2076
PRT / Med(I)	All	0.0023	0.0966	0.0238	0.3858
SRT	Manufacturing	-0.0042	0.1042	-0.0403	0.0614
PRT	Manufacturing	-0.0176	0.1113	-0.1581	-0.5533
SRT	Services	-0.0074	0.1138	-0.065	-0.0444
PRT	Services	-0.0119	0.1232	-0.0966	-0.1912

Table 4.1: Performance of hypothetical long-short portfolios

tor; regression (2) adds Mkt-RF to regression (1); regression (3) adds SMB, HML from Fama and French (1992) and momentum from Jegadeesh and Titman (1993) to regression (2). The risk-free rate is derived from the market yield on US Treasury securities at 1-Year constant maturity.

Table 4.2 reports the estimated factor loadings and risk premia. Regression (3) is the most interesting as it includes the most number of factors. All factors except for HML are statistically significant. Risk premium for the trade secret factor is 0.00109 with a factor loading of -0.328. In other words, trade secret risk has a positive correlation with expected asset returns. This is consistent with the capital asset pricing model which would suggest firms with high exposures to trade secret risks carry a risk premium over their peers.

4.3 Discussion

Portfolio sorting and the Fama-MacBeth regression are both commonly used asset pricing models, but why do they present contradictory results? The answer is neither of them - at least in their canonical forms - is ideal for testing the trade secret risk factor. To start with, portfolio sorting relies on factor scores and factor scores only to group assets into different sub-portfolios. Such a procedure is suitable for factors such as size or value because the market capitalization and book-to-market ratio are usually normally distributed among public companies. Trade secret risk, however, is far from being normally distributed. As shown in Figure 3.2, almost half of the firms have zeros as their trade secret risk factor scores. This means when sorting stocks into sub-portfolios, the bottom portfolio is over-crowded by firms with zero trade secret risks. Moreover, if each sub-portfolio has the same number of stocks then this over-crowding effect would "overflow" from the bottom to the middle portfolio as well. This problem can potentially be solved by using a normally-distributed factor to add a secondary sort after sorting on trade secret risks, but there is no clear answer to what this secondary factor is appropriate for this study.

	(1)		(2)		(3)	
Factors	Factor Loading	Risk Premium	Factor Loading	Risk Premium	Factor Loading	Risk Premium
Mkt-RF[1]			-2.814	-0.0325 (0.00722)	-0.435	-0.00428 (0.00138)
SMB					-0.118	-0.0108 (0.00192)
HML					-0.225	0.00112 (0.000646)
MOM					-0.168	-0.00996 (0.00141)
PRT	-0.785	-0.000182 (0.00026)	1.095	-0.000206 (0.000406)	-0.328	0.00109 (0.000301)
N	6706		6706		6706	
T	20		20		20	

Table 4.2: Coefficient estimates for the Fama-MacBeth regression

Fama-MacBeth regression avoids this problem by regressing returns on factors for each firm to get idiosyncratic factor loadings ($\hat{\beta}$s). Yet, a large T is required for those factor loading estimates to be accurate. In fact, most of the existing literature use monthly and sometimes weekly data in Fama-MacBeth regressions. This then poses a restriction on data availability - factors such as size and value are free from this restriction because market capitalization and book-to-market ratio are recorded every day. Trade secret risk on the other hand is proxied using annual reports, which means it is estimated once per year only. To keep regressions consistent, factor loadings on other factors also need to be estimated using annual data. However, some factors are not captured accurately with low frequency data. Book-to-market ratio, for example, undergo substantial changes as news break and earnings reports release. Inaccurate factor loading estimates lead to inaccurate risk premia estimates ($\hat{\lambda}$s).

Apart from the low-frequency issue, Fama-MacBeth regressions in general are flawed in three ways. Error terms in the cross-section are assumed to be uncorrelated, which is almost never the case for asset returns. If one company gets lucky, it is quite possible a similar company gets lucky as well. Petersen (2006) estimated that 41% of the literature simply ignores this violation, which causes the standard errors to be off by a factor of 10. In addition, Fama-MacBeth regressions do not account for changes in time. Variations in risk premia come solely from cross-sections, and any time-series variation is ignored. In short, while the estimates in Table 4.2 are statistically significant, the robustness are questionable.

5 Conclusion

In this study, I use natural language processing with company annual reports to produce a novel dataset that measures trade secret risks. I also test this risk factor with two asset pricing models, namely portfolio sorting and Fama-MacBeth regression. While portfolio sorting does not suggest trade secret risk as an effective factor, Fama-MacBeth regression shows a weak positive correlation between asset returns and trade secret risks. The result from the Fama-

MacBeth regression is consistent with the capital asset pricing model which suggests firms with high exposure to trade secret risks carry a risk premium.

However, the set-ups in both portfolio sorting and Fama-MacBeth regression are not ideal in some ways. Future research could address these issues in a few ways. For portfolio sorting, adding a reasonable secondary factor sort is an option. For Fama-MacBeth regression, trade secret risks can be captured and updated more frequently from quarterly reports and earnings call transcripts in addition to annual reports. More robust estimators can be produced with techniques such as clustering and Newy-West estimators.

References

Almeling, D. S. (2012). Seven reasons why trade secrets are increasingly important. *Berkeley Technology Law Journal 27*(2), 1091–1117.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2013, Aug). Quality minus junk. *SSRN Electronic Journal.*

Ben-Rephael, A., B. I. Carlin, Z. Da, and R. D. Israelsen (2021). Information consumption and asset pricing. *The Journal of Finance 76*(1), 357–394.

Black, F. (1972, Jul). Capital market equilibrium with restricted borrowing. *The Journal of Business 45*(3), 444–455.

Callen, J. L., X. Fang, and W. Zhang (2020). Protection of proprietary information and financial reporting opacity: Evidence from a natural experiment. *Journal of Corporate Finance 64,* 101641.

Campbell, J. Y. and R. J. Shiller (1988). Stock prices, earnings, and expected dividends. *The Journal of Finance 43*(3), 661–676.

Campbell, J. Y. and R. J. Shiller (1998). Valuation ratios and the long-run stock market outlook. *The Journal of Portfolio Management 24*(2), 11–26.

Castellaneta, F., R. Conti, F. M. Veloso, and C. A. Kemeny (2016). The effect of trade secret legal protection on venture capital investments: Evidence from the inevitable disclosure doctrine. *Journal of Business Venturing 31*(5), 524–541.

Cohen, L., C. Malloy, and Q. Nguyen (2020). Lazy prices. *The Journal of Finance 75*(3), 1371–1415.

Ettredge, M., F. Guo, and Y. Li (2017). Trade secrets and cybersecurity breaches. *SSRN Electronic Journal.*

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *The Journal of Finance 47*(2), 427–465.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Garcia, D. and O. Norli (2012). Geographic dispersion and stock returns. *SSRN Electronic Journal.*

Hirshleifer, D. A., P.-H. Hsu, and D. Li (2012, Dec). Innovative efficiency and stock returns. *SSRN Electronic Journal.*

Hsu, P.-H. (2009). Technological innovations and aggregate risk premiums. *Journal of Financial Economics 94*(2), 264–279.

Jamilov, R., H. Rey, and A. Tahoun (2021). The anatomy of cyber risk. *SSRN Electronic Journal*.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance 48*(1), 65–91.

Klasa, S., H. Ortiz-Molina, M. Serfling, and S. Srinivasan (2018). Protection of trade secrets and capital structure decisions. *Journal of Financial Economics 128*(2), 266–286.

Petersen, M. A. (2006). Estimating standard errors in finance panel data sets: Comparing approaches. *SSRN Electronic Journal*.

Png, I. P. (2017). Law and innovation: Evidence from state trade secrets laws. *Review of Economics and Statistics 99*(1), 167–179.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance 19*(3), 425.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review 71*(3), 289–315.