Using Representation Learning and Website Text to Identify Competitor Networks

Gerard Hoberg, Craig Knoblock, Gordon Phillips, Jay Pujara, Zhiqiang Qiu and Louiqa Raschid

This paper introduces a new approach to identify competitors using company websites to map competitive relationships among public and private firms. We apply representation learning techniques to create embeddings of companies based on website content, emphasizing information about industry-specific products and services. By placing all firms - public and private - within the same embedding space, we construct a peer competitor network that supports analysis of company interactions and market evolution. We label this new approach and set of network competitors as the Website Text-based Network Industry Classification (WTNIC). We evaluate the quality of the network using multiple ground truth datasets and benchmarks and show that it offers a robust alternative for understanding competition at scale. Despite using noisy website data, WTNIC matches or outperforms prior approaches that rely on curated data. We demonstrate the value of WTNIC through a case study examining how competition from China affects the public and private peers of a focal firm, highlighting the importance of placing both public and private companies within the same embedding space.

Key words: Competitor identification; Company embedding; Large Language Models; Doc2Vec; Representation learning; Peer networks.

Acknowledgments
We thank the National Science Foundation (grants 1561068 and 1937153), the Dartmouth Tuck School of Business, the USC Marshall Institute for Outlier Research in Business (iORB) and the Robert H. Smith School of Business at the University of Maryland for supporting this research.
The authors alone are responsible for any errors and omissions.

1. Introduction

Understanding the competitive structure of markets is important to researchers, investors, entrepreneurs, and government regulators alike. Economic and financial researchers have long sought to characterize competition (Hotelling (1929), Chamberlin (1933)). Empirical data-driven models of competitors within a sector are most useful when they cover the entirety of the market participants. Unfortunately, most readily available data and models only cover companies that are publicly listed on stock exchanges, whereas public companies are a small segment of the market comprising both public and private companies.

Our goal is to construct a peer competitor network that covers a large fraction of public and private companies within the same spatial embedding space. Private companies are widely under-researched due to data limitations, despite the value that such research can bring to financial markets. Identifying and modeling emerging private competitors will help to understand competitive threats and strategic opportunities for public companies. Going further, studies focused on private firm dynamics can better inform entrepreneurs, venture capitalists and policymakers.

Earlier approaches that identify competitors have relied on industry classifications (Pearce (1957)), financial metrics (Fama and French (1997), Bhojraj and Lee (2002)), job postings (Li and Wu (2025)), EDGAR search patterns for SEC 10-K filings (Lee et al. (2015)), newspaper articles and events (Hilt and Schwenkler (2020)), and self-reported peers (Rauh and Sufi (2011)). Industry classification also often required substantial manual effort by domain experts. This paper extends and improves upon prior approaches. Notably, it addresses the major limitation – a lack of training data for private companies – by using the vast and untapped resource of company websites, which cover both public and private companies. We label our peer network as the Website Text-based Network Industry Classification (WTNIC).

Our objective is to ensure that the network has a *primary focus on content about industry-specific products/services*. Further, the network will be constructed each year, allowing researchers to model the temporal evolution of competitors. By placing the public and private companies within the same embedding space, the peer competitor network will facilitate research seeking to examine the joint evolution of public and private companies, as well as the influence of private competitors on public companies.

Prior language-based approaches to construct competitor networks, e.g., TNIC (Hoberg and Phillips (2016)), relied on training data such as the SEC 10-K filings. The 10-K filings include an explicit section describing the products and services of the focal company and its competitors. Unfortunately, there is no similar corpus that covers private companies. Fortunately, website text does include content describing industry specific products and services. However, the website text is noisy and it also includes additional content, e.g., pages to establish a brand for the company and / or to engage with clients, pages targeting future employees, pages on corporate governance and ESG, pages to engage with investors, etc. Thus, a significant contribution of our research is an approach to develop a peer competitor network that targets only the content describing industry-specific products and services, while discounting additional content.

We construct a high-dimensional (spatial) vector embedding using unstructured website text that covers both public and private companies.[1] The competitor network is then constructed using the placement of each company within this embedding space, under the assumption that the embedding captures the nuances of competition. The spatial distance (closeness in the embedding space) between the corresponding pair of vectors is expected to be a proxy for the similarity between any pair of public companies, or the similarity between a private company and a public peer. In this context, the spatial distance between two companies that offer the same product or service

[1] This paper describes the construction and evaluation of a website text-based network industry classification (WTNIC) database, spanning the years 2000 to 2022, and including approximately 500,000 public and private companies each year. The WTNIC database will be made available under an appropriate license for free download. WTNIC has been supported by the National Science Foundation.

in the same market, i.e., direct competitors, should be small. In contrast, two companies offering unrelated products should have a large spatial distance between them.

Websites are a natural data source since companies have a vested interest in providing website text that targets customers and accurately describes their products and services. Website text is available for an extensive set of public and private companies over time, facilitating broad potential impact. We use representation learning techniques over website text to create a vector embedding space that captures peer relationships. We discuss a range of representation learning methods, including Large Language Models (LLMs), and present their advantages and disadvantages. We begin by creating a corpus of company website text using the Internet Archive's Wayback Machine. In addition to covering both public and private companies, our corpus will also have a temporal dimension that will span the time period of the archive. This has the potential advantage of modeling how the network of competitors changes over time.

While website text is widely available, extracting the product and service relevant content poses multiple challenges. The first challenge is that website pages can have diverse structures within and across industries. Landing pages can vary in their objectives, which can range from organizing geographic sites, to describing subsidiary organizations, to linking to product catalogs, and to focusing on corporate values such as ESG policies. Website text from different industries might also use similar terms but with different semantics. For example, the term `process` indicates different semantic intent in the legal industry and in manufacturing. In addition, industry vocabularies are not static, as disruptive companies appear and use new terms such as `ridesharing` or `wearables`, which must be seamlessly integrated into the vocabulary.

To illustrate the severity of this issue, an evaluation of an untuned competitor network based on unfiltered web text reveals that it incorrectly identifies large numbers of peers that are not even in the same broad industry sector as the focal company. These peers are very unlikely to have products and services similar to those of the focal company. This test conservatively defines sectors based on the very coarse Fama French 5 industry classification (Fama and French (1997)), making these findings quite stark. A main contribution of our study is the ability to fine-tune the learned representations such that selected peers actually have similar products and are mostly in the same industry sector.

The second challenge is that even when website text can be collected successfully, computing the embedding-based similarities between hundreds of thousands of companies can pose a formidable big data scalability bottleneck. For example, a pairwise network with 500,000 public and private companies, as used in this project, could require hundreds of billions of pairwise similarity calculations for a single year. We note that many similarity scores will be close to zero.

A key aspect of our research is the validation of the quality of the WTNIC network. To do so, we develop an extensive validation protocol that relies on multiple distinct ground truth datasets of competitors, drawn at varying granularity levels. This includes self-reported competitors and a novel ground truth database that is LLM+human-generated. Next, we demonstrate the financial and economic benefits of the WTNIC network beyond current competitor networks. We identify a set of financial metrics explored in corporate finance and asset pricing, and we examine how well the WTNIC peers can predict these metrics for the focal firm. To complete the evaluation, we utilize WTNIC to investigate the impact of a large shock that can affect broad industry sectors, e.g., competition due to the low cost entry of products from China. We explore the impact on both the public and private peers of a focal company.

Our research makes the following contributions:

- We demonstrate that the learned vector representation-based approaches are remarkably effective. Our proposed best WTNIC solution is built on a Doc2Vec embedding (Mikolov et al. (2013)) over website text that is further tuned so that the peer network focuses on products and services.
- The successful identification of competitors is the ultimate test of the effectiveness of any method. The evaluation compares peer competitors identified by our network to sets of

"ground truth" competitors from other data sources. WTNIC D2V+C500 outperforms the existing TNIC benchmark on these competitor ground truth tests. This confirms that the WTNIC website text embedding provides a strong platform for constructing a competitor network for both public and private companies.

- We complete an extensive validation of the WTNIC D2V+C500 peer network using multiple ground truth datasets. We complement the validation with two sets of contributions from a financial/economic perspective. First, we examine how competitors can predict the financial metrics of the focal firm. Second, we report results from an experiment to determine the impact of shocks from China on the public and private peers of a given focal public company.

The specific validation tests of our competitor network include the following:

- We confirm that the peers of our best WTNIC competitor network D2V+C500 are more likely to be In-Sector (IS) peers, defined as peers in the same Fama French 5 sector. Our tuned approach D2V+C500 outperforms all other competitor networks. For some Fama French 5 industry sectors, the ability to identify IS peers was exceptional. Our results are strongest for the `Health+` sector.
- Similarly, the tuned WTNIC D2V+C500 network shows very good performance at identifying peers in the same NAICS sector as the focal company, when considering 2-digit to 6-digit NAICS codes. A notable result is that D2V+C500 can predict NAICS classification codes for private companies (provided by Orbis) with high accuracy. This finding illustrates the quality of our (private, public) competitor network. We note that we are the first to make such predictions outside the narrow realm of public companies.
- We compare the overlap of D2V+C500 and the high-quality ETNIC network constructed using SEC 10-K filings (Hoberg and Phillips (2025)). This validation further attests to the high quality of the WTNIC embedding / D2V+C500 despite it being constructed using noisy website text.
- Finally, we introduce a novel and critical ground truth test based on a set of *LLM+human-generated ground truth peers*. For a set of randonly sampled private companies, we ask a Claude LLM instance to identify those publicly traded companies that are the most similar product market peers. We compute the recall (fraction of overlap) between the WTNIC D2V+C500 predicted peers and the LLM identified peers. We note that the fine-tuned WTNIC D2V+C500 network performs very well on this task.

Beyond ground truth tests of the peer network, we conduct financial and economic tests. We explore the ability of D2V+C500 peers to explain the following important financial variables commonly used in corporate finance and asset pricing research: 1. Net Operating Income over Total Assets (NOI/Assets or RNOA); 2. Net Operating Income over Sales (Profit); 3. Market-to-Book (M/B) Ratio; 4. Market Leverage; 5. Book Leverage; 6. HML Beta; 7. SMB Beta. The results suggest that WTNIC D2V+C500 peers exhibit a strong signal regarding their ability to explain focal firm variables.

We also report results from an experiment to determine the impact of competition from China on the public and private peers of a focal company. Shocks from China are large impact shocks that affect entire industry sectors, given the low-cost entry of products from China. The WTNIC embedding and peer network allows us to explore the impact of these shocks for both public and private companies. We note that little is known regarding how these shocks impact smaller private companies, in particular with regard to how they re-position themselves in the product market. This exploration thus underscores a primary strength of our website text based WTNIC network that includes both public and private peers.

The paper is organized as follows: Section2 summarizes research on competitor peer networks and representation learning. Section3 presents our methodological approach including (i) Representation learning; (ii) Tuning the vector embedding; (iii) Generating the competitor network. Section 4 summarizes the dataset and ground truth and explains our evaluation protocols. Section 5 presents the empirical results of an extensive evaluation. Section 6 presents the results of an experiment to determine the impact of competition from China on the public and private peers of a focal company. Section 7 concludes and discusses future research.

2. Related Work

We first summarize the literature on competitor identification. This research has its origins in the economics and finance literature. More recent activity has been reported in the Information Systems research domain, as companies expanded their website footprints, and provided new opportunities and datasets to study competition. Competitor networks are most useful when they cover the entirety of the market participants. Unfortunately, the majority of models rely on proprietary data, and are limited to public companies (a surprisingly small segment of the overall competitive landscape). We then briefly summarize research on representation learning and Large Language Models (LLMs) in finance.

2.1. Competitor Identification

Research has expanded beyond government SIC or NAICS classifications to identify competitors based on financial metrics such as asset pricing (Fama and French (1997)), asset holdings (Rauh and Sufi (2011)), valuations (Bhojraj and Lee (2002), Bhojraj et al. (2003)), or the co-movement of stock prices (Yaros and Imieliński (2015)). The underlying premise is that a latent industry assignment can play a role in predicting financial outcomes. An inferential process can then potentially estimate industry membership. A primary drawback is that they rely on financial data, which is difficult to obtain for private companies.

Other approaches use information from regulatory filings. The Text-Based Network Industry Classification (TNIC) (Hoberg and Phillips (2010, 2016)) used the co-occurrence of product / service specific terms in the SEC 10-K regulatory filings of public companies. Hoberg and Phillips (2025) include company scope and generate an improved version ETNIC - embeddings-based TNIC - which is based on Doc2Vec embeddings. The network is trained for each year so as to ensure the absence of any look-ahead bias. The primary drawback of ETNIC is its restriction to public companies.[2].

The explosion of online information about companies provides significant opportunities to explore competition further. Online sources range from company website pages (Pant and Sheng (2015)), to financial news / articles (Bernstein et al. (2002), Ma et al. (2011)), to website forums (Xu et al. (2011)), to large textual corpora including website search logs (Bao et al. (2008), Wei et al. (2016)). Research in (Pant and Sheng (2015)), in particular, examines competitor identification using corporate websites and website footprints. They propose several content and network based metrics as follows: 1. The similarity of inlinks (outlinks) to / from other websites; 2. The similarity of website text based on a well-known measure of document similarity; 3. The co-occurrence of a pair of companies in news articles or in search engine queries. We note that this approach also has the limitation that it does not address the task of building a competitor network for private companies. Some additional limitations are that news articles and online forums predominantly cover large and highly capitalized companies. Our approach of relying on website text has practical scaling advantages over the more labor-intensive data used in previous studies.

2.2. Representation Learning

Data-driven machine learning often uses representation learning to extract salient features from an underlying collection. When the collection is a set of documents, there is a history of drawing upon insights from natural language processing. The methods range from traditional, token-based approaches that represent words in a document, to distributional approaches that develop probabilistic models of language patterns. For example, word distributions over topics are used in the Latent Dirichlet allocation (LDA) Topic Model (Blei et al. (2003)). Extensions include supervised topic models (Blei and McAuliffe (2007)) and structured topic models (Roberts et al. (2013)).

[2] The website containing this data, https://hobergphillips.tuck.dartmouth.edu/, has attracted over 55,000 visits from academic and practitioner users from all over the world since its launch in 2010.

More sophisticated representation learning methods are better able to identify nuanced features from word sequences. For example, Doc2Vec embeddings (Mikolov et al. (2013)) can handle word sequences, while RoBERTa (Liu et al. (2019)), an efficient extension of BERT (Devlin et al. (2019)) can handle sentences and phrases. A summary of the state of the art for financial text mining using Large Language Models (LLMs) is presented in (Lee et al. (2024)). LLMs such as FinBERT-19, FinBERT-20 and FinBERT-21 (Araci (2019), Yang et al. (2020), Liu et al. (2021)) typically address general NLP tasks such as sentiment analysis, Named Entity Recognition, text summaries, text labeling, and question answering, in the finance domain.

Our approach using representation learning will be presented in the next section. We considered several alternatives including the simpler Doc2Vec embedding, as well as a RoBERTa model which was further trained using SEC 10-K filings. Our approach follows a trend in finance toward using more sophisticated models as they become available, starting with the use of Doc2Vec (Cong et al. (2024)), the use of BERT and RoBERTa (Hoberg and Philips (2024), Kolbel et al. (2024), and Acikalin et al. (2022)) ChatGPT (Jha et al. (2024) and Lopez Lira and Tang (2024)) and hybrid approaches combining embeddings with topic models (Hanley and Hoberg (2019) and Cong et al. (2024)). Like these novel hybrid approaches, our approach in this paper starts with embeddings and uses interpretable content to guide a fine-tuning transformation to a more informative spatial model.

We note that the more sophisticated approaches like RoBERTa have the limitation of requiring massive data collection(s) for training. However, given our tests in this paper that find lower performance with the RoBERTa model, it is unclear whether other LLMs can improve on the performance of Doc2Vec for website text. This is because website text can be noisy and often contains chunks of text that do not resemble the extensive training documents required by the more sophisticated models. We discuss this trade-off further in the paper, when we present our approach and evaluation results.

3. Methodology

Our approach to creating a competitor network introduces a multi-stage pipeline as illustrated in Figure 1. The first stage of the pipeline processes website text and applies one of several **Representation Learning** techniques to learn a high-dimensional vector for each company. The second stage of the pipeline is **Tuning** the vector embedding to increase the proportion of In-Sector (IS) peers. The third stage uses the learned vector representations to generate the set of potential competitors or the **Competitor Network**. The final stage of the pipeline, **Evaluation**, produces a set of tangible, economically-grounded predictions such as predicting financial metrics for a public company, or predicting the NAICS code for a private company, or predicting a set of peers identified using a generative AI tool.

The first stage of the pipeline processes website text and applies one of several **Representation Learning** techniques to learn a high-dimensional vector for each company. The second stage of the pipeline is **Tuning** the vector embedding to increase the proportion of In-Sector (IS) peers. The third stage uses the learned vector representations to generate the set of potential competitors or the **Competitor Network**. The final **Evaluation** stage of the pipeline conducts a set of validations. This includes the prediction of financial metrics for public companies, the NAICS code for a private company, or predicting self reported or AI generated peers for both public and private companies.

3.1. Problem Definition

Given a set of entities, E, the goal is to predict a set of competitor links, $L(e_i, e_j), \text{s.t.} e_i, e_j \in E$. In this paper, we assume that each entity e_i, has an associated corpus of documents, D_{e_i}. Our approach is to use a representation learning algorithm, Repr, and the corpus, to learn a k-dimensional representation, $R_{e_i} \in \mathbb{R}^k$, such that $R_{e_i} = \text{Repr}(D_{e_i})$.

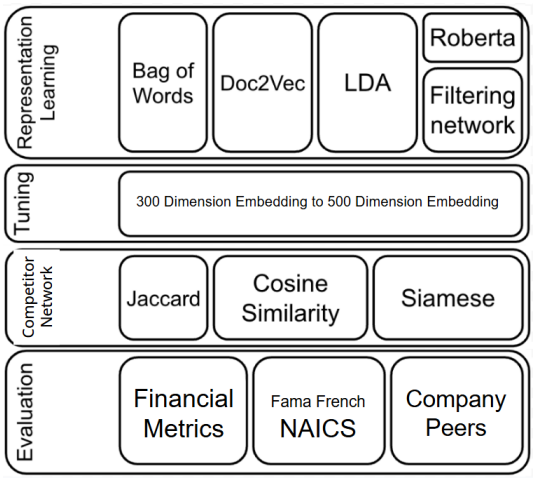

Figure 1 Architecture of the Pipeline to Create the Peer Network and Evaluation.

Next, these k-dimensional representations are used as input to a competitor link prediction algorithm, LP, to generate a set of pairwise scores between entities, $s(e_i, e_j) = LP(e_i, e_j)$. Finally, we apply a decision function, D, over the scores to determine the links to include in the competitor network: $\forall_{i,j} L(e_i, e_j)$ iff $D(s(e_i, e_j))$.

3.2. Representation Learning

In this paper, we experiment with several well-established approaches for representation learning. These include baseline models that use a bag-of-words representation, a probabilistic approach using latent Dirichlet allocation (LDA), and two deep learning approaches, Doc2Vec and RoBERTa.

3.2.1. Bag of Words: Prior work (Hoberg and Phillips (2016)) represented companies by the words that occurred in the business description section of the annual SEC 10-K filings. Filters were applied to remove high-frequency terms and retain only nouns and proper nouns. We do not use any additional filtering, e.g., parts of speech. Each company will be represented by a vector of word occurrence, and these vectors will be normalized to have unit length.

3.2.2. Latent Dirichlet Allocation: Topic modeling is a popular approach to identifying the latent feature space, or topics, with a large document corpus. Topic models typically use unsupervised probabilistic models to associate individual documents with a probability distribution over latent groups, known as topics. Latent Dirichlet allocation (LDA) (Blei et al. (2003)) is a sophisticated probabilistic model that has served as the foundation for much recent work in topic modeling. The method, defined for k topics and $|D|$ documents, models topics as distributions over words (β_k), and document as mixtures of topics (θ_d). For each word $w_{d,n}$ in a document $d \in D$, a topic $z_{d,n}$ is chosen from the document's topic distribution (θ_d), and a word is chosen from the topic's distribution over words (β_k). The Dirichlet distribution of θ_d and β_k are parameterized by terms α and λ respectively. This generative approach is reflected in the joint distribution shown in Equation 1. Topic models are learned through inference on the generative model described. This inference problem is intractable at scale, and approximate techniques such as Gibbs Sampling or variational methods are used during learning (Blei et al. (2003)).

$$p(w, z, \theta, \beta | \alpha, \lambda) = \prod_k p(\beta_k|\lambda) \prod_d p(\theta_d|\alpha) \prod_n p(z_{d,n}) p(w_{d,n}|\beta_{z_{d,n}}) \tag{1}$$

To use topic models for representation learning, we use the inferred topics for a given document as the learned representation. Since each company website is associated with many website pages, we concatenate the content of these pages into an aggregate document and we learn a distribution

over the aggregated document. Details of parameters for model training are in the evaluation section.

3.2.3. Doc2Vec: Word embeddings represent each word as a vector of a m-dimensional vector, most commonly in $\mathbb{R}^m$. Many approaches have been proposed for training word embeddings, spanning from early work in latent semantic indexing (Hofmann (2017)), to more recent tools such as Word2Vec (Mikolov et al. (2013)) and GloVe (Pennington et al. (2014)). We adopt a variant of the Word2Vec model, Doc2Vec (Le and Mikolov (2014)). Doc2Vec differs from Word2Vec in that it uses a global contextual vector associated with the document as part of its predictive model.

The training objective of Doc2Vec is to learn word vector representations that are good at predicting nearby words. In the model, $v(w) \in \mathbb{R}^m$ is the vector representation of the word $w \in W$, where W is the vocabulary and m is the embedding dimensionality. $v(e_i) \in \mathbb{R}^m$ represents a latent vector associated with each document. Given a pair of words (w_t, c), the probability that the word c is observed in the context of word w_t in a document associated with a latent vector e_i is as follows:

$$P(c=1|v(w_t), v(c), v(e_i)) = \frac{1}{1+e^{-v(wt,e_i)^T v(c)}}$$

Given a training set containing the sequence of words, the word and document embeddings are learned by maximizing the log-likelihood score. For the collection of website pages for each company, we learn a single latent document vector for the company, as well as the vectors for each of the individual words across all the pages. Through this process, each company will be associated with a vector that models the pattern of contextual word co-occurrence across the website pages for that company.

3.2.4. RoBERTa: The final representation learning approach we apply is RoBERTa, a contextual, transformer-based pre-trained language model. In contrast to Doc2Vec, which is primarily trained to learn word representations from a fixed set of contextual words, RoBERTa is designed to learn representations based on a sequence of contextual tokens, resulting in a different representation for a word based on its surrounding context. Using RoBERTa, we map a sequence of words, $w_1, w_2, ..., w_n$ to a vector in $\mathbb{R}^m$. A limitation of RoBERTa is that the sequence of tokens for which it can generate representations is limited to 512 tokens. Most websites will yield many chunks of 512 tokens, and the diversity of chunks makes generating a combined representation difficult. Alternatives such as LongFormer (Beltagy et al. (2020)) have been proposed to overcome this limitation; however, such approaches require substantially more computational resources.

To overcome the limitation of 512 tokens, we develop a filtering technique to identify the most relevant portions of a company website. To do this, we train a classifier to differentiate between text that characterizes a company's product or services versus less relevant text. As positive examples of text, we use the first paragraph of the business description section in SEC 10-K filings. As negative examples, we use documents from an NLTK standard corpora (Loper and Bird (2002)), e.g., the Brown Corpus. We introduce a filter head (Figure 2), which is trained to classify these documents into positive and negative classes. Using the RoBERTa representations from the CLS tag of each document, we train a fully connected multi-layer Perceptron to predict the class of each document.

We incorporate the trained filter head and classifier in the RoBERTa model as shown in Figure 3. Each document is first split into smaller chunks of 512 tokens. These chunks are each assigned a probability of being relevant to the company's business function. We rank these chunks and utilize the CLS embeddings of the top K chunks for each company.

Let $\vec{Z^j}$ be the output vector of the filter head for chunk j, $\vec{Z^j} = f(c_j) = [Z_+^j, Z_-^j]$. Let S_+ and S_- be the (softmax) scores to predict that the chunk is business relevant (+) or general text (-). We define the relevance score for chunk j, denoted as RS_j, as follows:

$$RS_j = \frac{e^{Z_+^j}}{e^{Z_+^j} + e^{Z_-^j}}$$

To maintain an equitable representation across documents, we use RS_j to order all token chunks in the document, and select the top K scoring chunks. These chunks are combined by computing the arithmetic mean over the k chunks, producing a vector in $\mathbb{R}^m$.

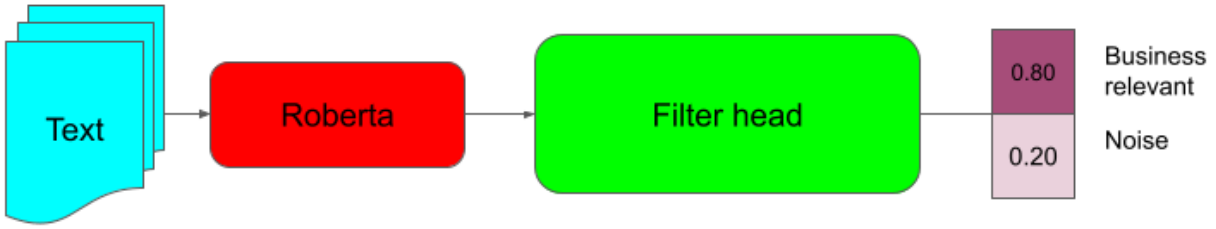

Figure 2 **Filtering technique to classify relevant chunks of a long document for use in RoBERTa.**

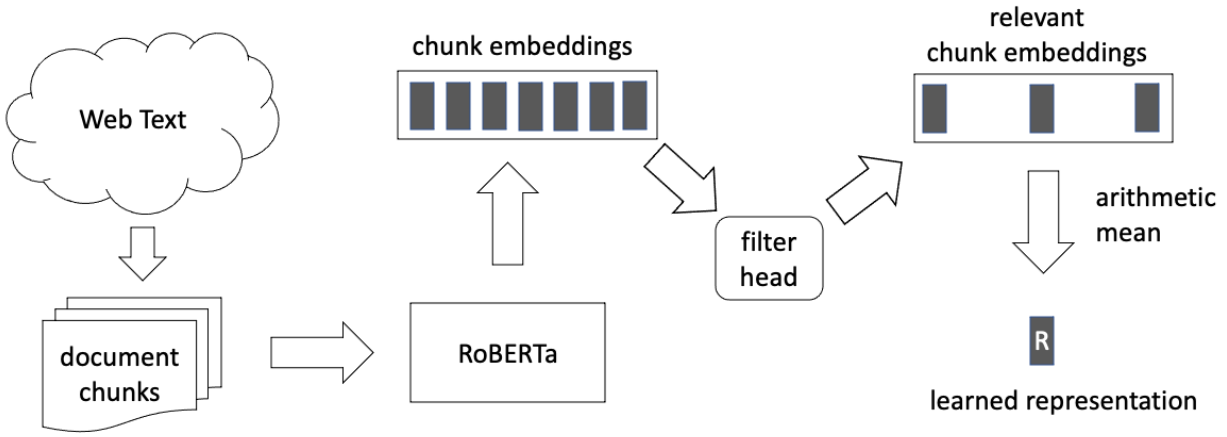

Figure 3 **Protocol to create a single representation R from the chunks of a long document using RoBERTa.**

3.3. Tuning the Doc2Vec Vector Embedding: D2V+C500

Our tuning approach begins with the ETNIC embedding over 10-K filings for public companies (Hoberg and Phillips (2025)). We use this ETNIC embedding to identify 500 industry clusters, labeled C500, with each cluster representing a subset of public companies that are strongly associated with an industry specific product or service domain. Our tuning approach will use these C500 clusters to compute a 300 X 500 tuning matrix. We can use this tuning matrix to transform the 300 dimension vector, representing each company in the original Doc2Vec embedding over website text, into a *500 dimension tuned vector in the D2V+C500 embedding space.*

The 500 clusters and the 300 X 500 tuning matrix will be computed for each year. We refer to this method as the D2V+C500 method, and the resulting competitor identification as the Webpage Text-Based Network Industry Classification (WTNIC). The tuning is a linear transformation from the 300-dimension Doc2Vec embedding space to a 500-dimension embedding space. As a result, creating the peer networks is scalable in the context of potentially hundreds of thousands of private companies. The breakthrough for our D2V+C500 method is that the resulting tuned 500-dimensional embedding indeed differentiates between industry-specific content about products/services and irrelevant or boilerplate text.

As is well known in the machine learning and Large Language Model (LLM) literature, it is a very difficult task to fine-tune an embedding without requiring significant amounts of training

data. Despite this limitation, the resulting peer network D2V+C500 performs very well for both public and private companies in our ex post validation tests.

The specific tuning protocol is as follows:

1. We use the ETNIC embedding over 10-K filings for public companies (Hoberg and Phillips (2025)) to perform a K-means clustering, and to identify 500 industry clusters, labeled C500. Each cluster will include a subset of public companies that are strongly associated with an industry specific product or service domain. An extra step is taken to ensure that each cluster has a minimum of three companies. The choice of 500 clusters balances the need for high granularity to capture a complete set of product markets with the need to also ensure adequate critical mass in each cluster.
2. Next, for each of the 500 clusters, we run a fitting regression in which each company is an observation. We consider the 300 dimension vector(s), representing the company in the original Doc2Vec embedding space, as the independent variables. For each company observation, we define a dummy dependent variable (outcome) to be equal to "1" if the company has been assigned to that cluster; the outcome is set to be a "0" otherwise. We perform an Ordinary Least Squares regression over the company data points. There is no multicollinearity as the 300 dimensions in the Doc2Vec embedding space are approximately orthogonal. The coefficients produced in this step are used to form a 300 dimension vector, one for each of the 500 industry specific clusters in C500. This will result in a 300 X 500 coefficient based tuning matrix.
3. In the final step, we use the 300 X 500 tuning matrix to perform a linear transformation, for each of the public or private companies, from the original 300 dimension Doc2Vec embedding space to the tuned 500 dimension D2V+C500 embedding space. Then for each pair of companies, we compute the pairwise cosine similarity using the pair of 500 dimension vectors, one for each company, to determine the similarity score for that pair.
4. This score will be used to construct the public-public peer competitor network of (public, public) peers, as well as the public-private network of (public, private) peers. We define industry peers as those firm-pairs having a similarity above a specified threshold. Following Hoberg and Phillips (2016), this threshold is chosen so that 1% or 2% of firm pairs are deemed to be industry peers.

3.4. Generating the Peer Competitor Network

The task of predicting a relationship between two entities is known as link prediction; we cast the problem of inferring the competitor network between companies as a link prediction task. For a given company entity, e_i, we use an associated set of documents (d_i) to learn a representation $R(d_i)$. Given learned representations, $R(d_i)$ and $R(d_j)$ for entities i and j, the goal is to assign a probability score $s(e_i, e_j)$ that the two entities have a competitive relationship.

There are two strategies for link prediction. The first strategy is an unsupervised approach: cosine similarity. The guiding premise is that documents containing similar information should have similar representations in the high-dimensional vector space. The second strategy is a supervised approach to learning a mapping from the learned representation to a competitor score. The advantage of unsupervised approaches is a lack of reliance on training data. Training data for understanding competition is scarce and the majority of training data is biased towards larger publicly traded companies. There are also many diverse definitions of competition. Thus, supervised approaches may be biased or suffer from a lack of training data.

Cosine similarity is an unsupervised approach and is known to be a reliable metric to measure the similarity of the embeddings in a high dimensional space. We use cosine similarity to calculate the pairwise similarity between the learned vector representations. We compute cosine similarity of the two learned representations $R(d_i)$ and $R(d_j)$ as:

$$sim_{i,j}^{cos} = \frac{R(d_i) \cdot R(d_j)}{||R(d_i)||||R(d_j)||}$$

The final step of our pipeline is generating the peer network. Our approach to determine which links to include uses a global scoring approach. We generate all pairwise company scores, normalize the scores for each focal company, and retain the top X% scoring links. This procedure ensures that the number of peers generated is consistent across methods since it produces a fixed granularity of peers. Further, it naturally adapts to variability in the peer scores for a focal company; a larger company may have many higher-scoring competitor links, while a smaller company may have fewer competitor links. We found that a threshold of the top 2% for the public-public network, and the top 1% for the larger private-public network, worked well and provided a competitor network that had similar properties compared to the TNIC network.

4. Evaluation Protocol

We consider the following evaluation tasks using multiple external metrics to compare the effectiveness of the learned representations and the WTNIC competitor network:

1. Financial metrics for public companies.
2. Self reported peers for public companies.
3. In-Sector (IS) peers based on the Fama French 5 sectors.
4. NAICS and SIC codes for private and public companies.
5. ETNIC competitors for public companies.
6. LLM+human generated competitors for private and public companies.

The 10-K is a comprehensive report filed annually by a publicly-traded company about its financial performance and it is required by the U.S. Securities and Exchange Commission (SEC). The 10-K was the primary source of ground truth for public companies. Proprietary databases from CapitalIQ and Orbis were the source of (limited) ground truth for private companies. In the rest of this section we provide details of the ground truth and the evaluation process. Then we describe the data pre-processing and experiment setup for these tasks.

4.1. Prediction Tasks and Ground Truth Labels

4.1.1. Prediction of Company Financial Metrics using WTNIC peers: Following the approach of (Bhojraj and Lee 2002, Bhojraj et al. 2003), we evaluate the effectiveness of our competitor relationships based on the ability to predict financial metrics using these competitors. We curated a subset of 3800 public companies from the SEC 10-K filings corpus for 2016, and report on the results of the evaluation tasks for this subset. We consider a range of variables explored in the corporate finance and asset pricing literature. The ability to predict these variables using competitors is a proxy of the quality of the peer networks. We consider the following metrics: 1. Net Operating Income over Total Assets (NOI/Assets or RNOA); 2. Net Operating Income over Sales (Profit); 3. Market-to-Book (M/B) Ratio; 4. Market Leverage; 5. Book Leverage; 6. HML Beta; 7. SMB Beta.

We compute these metrics using the competitors R_i for focal company c_i. The competitors are generated using the appropriate competitor network. We fit a regression model to estimate $\hat{F}$ based on the average value of that metric $\overline{F(R_i)}$, over all competitors, and we learn values for λ and c:

$$\hat{F}(c_i) = \lambda \overline{F(R_i)} + c$$

We then evaluate the quality of the set of competitor relationships on the basis of the coefficient of determination:

$$R^2 = 1 - \frac{\sum_i (F(c_i) - \hat{F}(c_i))^2}{\sum_i (F(c_i) - \overline{F})^2}$$

A high value for R^2 suggests that the regression over mean value for the metric, based on the competitor network, successfully explains the observed financial results for the focal company c_i.

We also consider the quality of the prediction for individual companies c_i by using the root mean square error (RMSE) as follows:

$$RMSE = \sqrt{\frac{\sum_i (F(c_i) - \hat{F}(c_i))^2}{N}}$$

We note that we report on two options for the peer network used in the regression. A peer network / regression that does not include the focal firm (self) as a peer reflects the pure predictive power of the peers. Regressions that do include self as a peer indicate the total level of explanatory power of the peers. This latter approach to include self resembles the norm for industry fixed effects models, where the focal firm is included in the respective industry group.

4.1.2. Ground Truth Self-reported Competitor Peers: Publicly traded companies typically self-report their competitors in the SEC 10-K annual filing. A company, e.g., a bank, may compete with all other companies in their sector,e.g., other banks. Typically, each company may strategically list only some of its true competitors, and it may also report some false competitors. Competitors may be identified and discussed in Item 1A (Risk Factors) or Item 7 (Management's Discussion and Analysis).

4.1.3. Ground Truth Competitors for In-Sector (IS) Peers: We evaluate each model regarding whether its peers are in the same broadly defined industry sector. This test is stark as a high quality model should choose fewer peers that are not even in the same sector. Following convention in the literature, we define sectors using the SIC code mappings from Fama and French (1997). For simplicity, we use a reduced grouping of the following five industry sectors to determine In-Sector (IS) peers:

- (Consumer) Durables and Non-Durables; Wholesale; Retail; Some Services.
- (Manufacturing); Energy; Utilities.
- (HiTec): Business Equipment; Telecommunications; TV; Control; Computational Services.
- (Health); Healthcare; Medical Equipment; Drugs.
- (Other) Multiple sub-sectors; Finance.

4.1.4. Ground Truth for NAICS and SIC Codes: NAICS is a six-digit hierarchical classification system from the US Census Bureau (Office of Management and Budget (2017)). The 2-digit prefix represents a set of 22 categories; each subsequent digit beyond the 2-digit prefix represents sub-categories. SIC is a 4-digit code (Pearce (1957)) from the US Department of Labor. Publicly traded companies include NAICS and SIC codes in their 10-K filings. Private companies do not have to self-report their industry classification. Several widely used databases such as CapitalIQ and Orbis report on NAICS and SIC codes for private companies. We will focus on private companies where the NAICS code has been validated by two sources or where it is obtained from a single high-quality source. We note that the US government wishes to adopt the NAICS code universally, but SIC codes are used more widely in industry for historical reasons. This push for widespread adoption increases the importance of identifying the NAICS sector for private companies.

We report on the percentage of peers who share the same NAICS code. A majority vote over the NAICS codes of these peer companies will then be used for prediction. We first consider the 2 digit prefix and continue until we reach the complete 6 digit NAICS code.

4.1.5. ETNIC Ground Truth competitors: Going beyond ground truth based on SIC or NAICS codes, we consider a more stringent set of ground-truth competitors. As noted earlier, ETNIC (Hoberg and Phillips (2025)) is a high-quality peer network that is based on an embedding constructed over the SEC 10-K filings. We compare the direct network overlap of Doc2Vec and D2V+C500 with ETNIC using threshold granularities ranging from 2% to 20% of the networks.

4.1.6. LLM Generated Ground Truth Competitors: We curate a novel set of *LLM-generated ground truth peers*. For each focal private company in a random sample, we ask a Claude LLM to identify those publicly traded companies that are its most similar product market peers. Details are provided in the paper. The ground truth dataset - ClaudeGT - represents the responses from the Claude LLM; it includes almost 7,500 entries referencing 660 private companies and almost 1,600 public companies. We note and account for the fact that generative AI models like Claude are biased towards large(r) public companies that are more prevalent in open training datasets.

4.1.7. Competitor Prediction Task: To evaluate the WTNIC set of competitors versus these different competitor ground truths above (4.1.2-4.1.5), we generate a competitor network from the full set of webpage similarities, based on a fixed network granularity of 2% or 1%. Our method generates an initial symmetric bidirectional network; however, the chosen subset of 2% or 1% may not be symmetric. In the real world, we expect many competitor relationships between firms to be bidirectional. However, based on the self-reported competitors, less than half of the competitor relationships in our ground truth are bidirectional. For this evaluation, we thus considered both a directed and bidirectional version of the ground truth relationships. We did not observe significant differences with respect to the different approaches. We report on the results from the bidirectional ground truth.

4.2. Data Pre-processing

The primary source for historical company website pages and website text is the Wayback Machine project of the Internet Archive. Our first corpus comprises the website pages of approximately five thousand public companies and 800,000 private companies, over a span of up to 20 years. We identified approximately 800,000 private companies and their URLs from the Orbis and CapitalIQ databases. We then crawled the Internet Archive using the URLs. The Internet Archive protocol for crawling follows a set of standards that produces high-quality results. For the limited number of cases where the crawl was not successful, we supplemented the corpus with additional pages from the company website.

A second corpus comprises the 10-K annual filings with the SEC, collected following the protocol in Hoberg and Phillips (2016). We report on results for the two corpora from 2016.

We apply the following pre-processing steps:

- SEC 10-K filings: We include words formed with letters from the English alphabet and excluded other text or characters, including numbers, punctuation, and special characters.
- Website text: The raw HTML files are processed using the Beautiful Soup library; this step removes tags, java code snippets, etc. We use all pages up to a depth of three from the root page. We replace special characters, characters not in the English alphabet, numeric digits not in 0-9, and punctuation with spaces. We replace longer whitespace and tabs with a single space. We retain periods(.), commas(,), and quotes (",').

We experimented with different document-cleaning processes. This included using space-separated text without any alterations to the raw text; replacing all punctuation and special characters with spaces; replacing only selective punctuation with space; leaving some special characters undisturbed. We included the English alphabet, numbers, and a few punctuation and special characters, e.g., period(.), comma(,), and quotes (",'). These variations produce comparable results with less than 1% variation.

From the SEC 10-K filings corpus for 2016, we curated a subset of 3800 public companies and report on the results of the evaluation tasks for this subset. The subset met the following conditions:

1. The company website text contained meaningful content.
2. The company reported at least one NAICS code.
3. The company reported profitability data.
4. At least one competitor was included.

For the private companies, we randomly sampled a subset of 32,000 companies from the corpus of approximately 800,000 companies. The sampling was performed so that the distribution of companies over the five Fama / French sectors was similar for both the public and private companies.

4.3. Parameters for Model Training

Doc2Vec: We use the Gensim Doc2Vec implementation for training the document embedding (Rehurek and Sojka (2011)). The best combination of hyperparameters was 100 epochs of training, the paragraph vector (PV-DBOW) setting, a dimensionality of 300, 1e−5 downsampling, and 10 negative sampling (Mikolov et al. (2013)). We observed that using negative sampling over hierarchical softmax (Mikolov et al. (2013)) performed better for the profit prediction task. We report on the model constructed from the 2016 dataset.

LDA: The key parameter to tune the LDA models is the vocabulary size of the corpus. We limited the vocabulary to only include words that occur in at least two documents; we also remove words that occur in more than 20% of the documents. We limit each document (company) to 2400 words. Documents are tokenized into a bag of words (BOW). The model was trained with the Gensim ldamulticore model with 100 topics and 200 passes. We also experimented with 500 topics when we used the topics to tune the Doc2Vec / Cosine similarity scores to produce the D2V+LDA peer network.

RoBERTa: We use the fairseq implementation of a pre-trained RoBERTa model (Ott et al. (2019)). For the SEC 10-K filing document corpus, we used the text from Item 1 (Business). We did not filter out any special characters and we used word chunks of 500 words.

To process the website text corpus for RoBERTa, we applied a trained filter head over the document content, and we filtered out only the top 10 most relevant word chunks of 500 words each. The protocol to train the filter was described in §3.2. We used 3943 instances of SEC filings (positive labels) and 3670 instances of general documents from the NLTK Brown corpus (negative labels). We used the Adam optimizer and the PyTorch platform with cross-entropy loss.We performed a five-fold cross-validation on the filter head to assess its accuracy. We were able to achieve an accuracy of 99.94 ± 0.08 showing that the filter head can easily differentiate between relevant documents discussing products and services and noisy or irrelevant documents.

5. Evaluation Results

5.1. Results Summary

Tables 1 and 2 report on an extensive evaluation of our approach to construct a high-quality fine tuned embedding. The evaluation is presented as a decomposition, with one dimension considering corpus quality, ranging from the high-quality SEC 10-K corpus for public companies to the very noisy Webtext for hundreds of thousands of private companies. We experiment with a range of methodologies along a second dimension, to construct the embedding over the corpora. Key findings are as follows:

- We demonstrate that learned representation-based approaches are effective for competitor identification. ETNIC, a Doc2Vec embedding applied to the 2016 10-K corpus for 3800 public companies (Hoberg and Phillips (2025)), improved on profitability prediction results from the original TNIC benchmark (Hoberg and Phillips (2016)).
- Competitor recall and ground truth tests (we provide details on ground truth tests later) are the key evaluation metric for a competitor network. The Doc2Vec embeddings that are trained on either (i) the website text for public companies, or (ii) the website text for both public and private companies, are successful at the task of recalling self-identified competitors.
- Our proposed solution, WTNIC D2V+C500 - a Doc2Vec embedding fine-tuned to focus on product and service content, yields substantial improvement in identifying peers in the same industry sector. This performance notably surpasses the baseline performance corresponding

to the TNIC and ETNIC benchmark networks, trained on the high quality SEC 10-K corpus. We view the good performance of D2V+C500 to be significant since the embedding covers both public and private companies.

Table **??** reports on the ability of D2V+C500 to predict several financial variables. The results suggest that website text-based peers have a strong signal regarding their ability to explain each of these variables. We find that the results do not change materially when adding controls for the company assets and age, indicating that the website text-based signal is both relevant and stable.

Fine tuning the peer network to focus on product-market peers is an important goal that motivates our next test that assesses each network's ability to identify In-Sector (IS) peers. Table 4 reports the results. Notably, D2V+C500 shows the best performance at identifying In-Sector (IS) peers as defined by the Fama French 5 industry sectors. The use of the very coarse Fama French 5 sectors makes this ground truth test stark. High-quality networks should not identify many peers that are not in the same coarsely defined industry sector. For some industry sectors, the ability of D2V+C500 to identify IS peers was exceptional, suggesting that companies in these sectors use more focused language on their websites.

Tables 5 and 6 report on the performance of D2V+C500 to predict ground truth peers defined based on NAICS codes for both focal public and private companies, respectively. We report results for 2-digit to 6-digit NAICS codes. The results indicate outstanding performance; we also emphasize that we are the first to attempt such ground truth predictions for private companies.

Table 7 extends our tests beyond ground truth based on SIC or NAICS codes as such industry classifications have limitations. We consider a more stringent competitor identification test based on ETNIC, a high-quality peer network that is based on embeddings constructed using SEC 10-K filings (Hoberg and Phillips (2025)). Table 7 compares the direct network overlap of Doc2Vec and D2V+C500 with ETNIC. Table 8 ends the evaluation with a novel and critical test based on a set of *LLM-generated ground truth peers* where we use Claude (specifically claude-3-5-sonnet-20240620) to produce a set of peers for private firms. We note that our fine-tuned Doc2Vec model D2V+C500 performs extremely well on this task as well.

5.2. Evaluation of Corpus Quality Versus Methodology to Construct an Embedding

Table 1 reports on the performance of a range of datasets and models. The evaluation is presented as a decomposition, with one dimension considering corpus quality. This ranges from the high-quality 10-K corpus for public companies (Panel 1) to the noisy Webtext for the public companies (Panel 2) and for tens of thousands of private companies (Panel 3). Within each panel, we experiment with a range of methodologies along a second dimension, to construct the embedding over the corpora. Thus, a comparison of Panel 1 versus Panels 2 and 3 can assess the impact of corpus quality versus methodology to construct an embedding. The evaluation metrics are reported for a profit prediction task and a competitor recall task, using the 2% threshold of peers for the focal company. All results are reported for the 3800 public companies. We report the adjusted average R^2 value, where the dependent variable is the Net Operating Income over Total Assets (NOI/Assets or RNOA), or the Net Operating Income over Sales (Profit) of the focal company. The independent variable is the value of that metric, averaged over the 2%-threshold peers for the model. Table2 reports on these metrics in greater detail.

Panel 1 reports on models constructed using multiple approaches over a high-quality corpus comprising the 2016 SEC 10-K filings. ETNIC (Hoberg and Phillips (2025)), a Doc2Vec learned representation using Cosine similarity to construct the competitor network, significantly outperforms the TNIC benchmark (Hoberg and Phillips (2016)). We note that TNIC uses a bag-of-words approach and Jaccard similarity without any sophisticated techniques for filtering and normalizing data. The recall for ground truth competitor prediction for ETNIC is 47.94% versus 31.39% for TNIC. The R^2 value for profit prediction increases from 0.3666 (TNIC) to 0.4750 (ETNIC = Doc2Vec/Cosine). A surprising result is that a more sophisticated approach, e.g., RoBERTa (Devlin et al. (2019))

Table 1 Profit Prediction and Competitor Prediction for Public Companies

Model	Similarity	Average Profit R-squared	Monopoly Rate(%)	Competitor Recall(%)
Panel 1: Multiple Approaches over the SEC 10-K filings for 3800 public companies				
TNIC	Jaccard	0.3666	17.63	31.39
BoW	Jaccard	0.3605	43.30	24.90
LDA	Cosine	0.3668	4.37	37.99
ETNIC (Doc2Vec)	**Cosine**	**0.453**	**0.026**	**55.87**
RoBERTa	Cosine	0.3456	19.81	33.76
Panel 2: Multiple Approaches over website text for 3800 public companies				
BoW	Jaccard	0.2497	34.20	13.53
LDA	Cosine	0.2858	0.89	19.88
Doc2Vec	**Cosine**	0.418	0.34	**30.20**
RoBERTa	Cosine	0.3455	13.60	11.10
Panel 3: Approaches over website text for thirty two thousand private companies				
LDA	Cosine	0.276	13.4	21.40
Doc2Vec	Cosine	0.403	0.42	30.57
D2V+C500 (Doc2Vec)	**Cosine**	0.348	6.4	**39.86**

The evaluation is presented as a decomposition. One dimension considers corpus quality, with the three panels ranging from the high-quality SEC 10-K corpus for public companies to the very noisy Webtext for hundreds of thousands of private companies in the third panel. Within each panel, we experiment with a range of models along a second dimension, to construct the embedding over the corpora. Each model is indicated using the first two columns. The third column displays the adjusted average R^2 (R-squared) of an OLS regression, where the dependent variable is the Net Operating Income over Total Assets (NOI/Assets or RNOA), or the Net Operating Income over Sales (Profit) of the focal company. The independent variable is the value of that metric, averaged over the 2%-threshold peers for the model. The monopoly rate is the fraction of companies that do not have any 2%-threshold peers; a low(er) value is better. Competitor recall is the percentage of self-declared ground truth competitors, and IS Peers is the percentage of peers in the same Fama French 5 sector as the focal company, that are included among the 2%-threshold peers.

does not appear to perform very well. It significantly underperforms ETNIC and is closer in performance to the simpler TNIC benchmark. We discuss this result further in the context of the next panels.

Panel 2 of Table 1 reports on models constructed over a corpus comprising the website text for the 3800 public companies. We observe that Doc2Vec using Cosine similarity continues to demonstrate good performance for this corpus. However, performance degrades as we go from Panel 1 to 2. In comparison to the 10-K-based ETNIC in the First Panel (47.9% recall), the recall for competitor prediction for Doc2Vec / Cosine is 30.2%. The Average R^2 for profit prediction also reduces from 0.4750 (ETNIC) to 0.418. These lower evaluation results reflect some expected degradation of signal quality when going from the fully curated and well-structured SEC 10-K corpus to the noisy website corpus. Overall, this indicates that website text captures sufficient signal, despite the noise, motivating our work to fine tune the website embeddings and to ultimately construct a rich competitor network comparable to those curated from the SEC 10-K corpus.

In both Panels 1 and 2, we note that RoBERTa (Devlin et al. (2019)) does not perform well. The RoBERTamodel used a pre-trained set of parameters that were not tuned for the website text dataset. Further, website text is often noisy and may not have normal sentence structure. In contrast, Doc2Vec has the advantage of being trained directly on the website corpus. These differences may explain the mediocre performance of the more sophisticated RoBERTamodel.

Panel 3 of Table 1 shows the performance as we expand the corpus to also include the website text for thirty-two thousand private companies. The companies were randomly sampled, from a collection of 800,000 private companies, to reflect a similar distribution of the 3800 public companies over the Fama French 5 sectors. Of particular interest is that the fine-tuned Doc2Vec method WTNIC D2V+C500 increases competitor recall to 40.3%, despite being constructed over noisy Webtext. This result is notable when compared to the recall of 47.9% of ETNIC that is constructed over the high quality corpus of SEC 10-K filings. We further note that the value of R^2 for D2V+C500 declines somewhat to 0.348. We justify this tradeoff since we believe that the metrics of competitor recall and In-Sector (IS) peers are the most appropriate metrics for evaluating a competitor network.

Table 2 IS Peers and Profit Prediction - Assets and Sales - for Public Companies

Model	Similarity	% IS peers		Profit/Assets		Profit/Sales		logage,logassets
		2%	1%	R-sq	RMSE	R-sq	RMSE	
Panel 1: Embedding over website text for 3800 public companies								
Doc2Vec	Cosine	66	73	**0.442**	0.197	**0.394**	0.303	
Doc2Vec	Cosine	66	73	**0.503**	0.186	**0.447**	0.290	yes
Panel 2: Embedding over website text for thirty two thousand private companies								
Doc2vec	Cosine	72	79	0.427	0.199	0.380	0.306	
Doc2vec	Cosine	72	79	0.492	0.188	0.436	0.292	yes
D2V+C500	Cosine	**92**	**97**	0.373	0.209	0.326	0.320	
D2V+C500	Cosine	**92**	**97**	0.469	0.193	0.412	0.299	yes
LDA	Cosine	68	80	0.298	0.220	0.254	0.336	
LDA	Cosine	68	80	0.432	0.198	0.375	0.308	yes

Each model is identified in the first column. The next two columns report on the fraction of peers, at either the 2%-threshold (used by HP2016), or the 1%-threshold, that are in the same sector as defined by the Fama-French-5 sectors; a higher percentage is better. The fourth and fifth columns display the adjusted R^2 and RMSE of OLS regressions where `Profit/Assets` *is the dependent variable and the average over the 2%-threshold peers is the RHS variable. The next two columns report similar results for* `Profit/Sales`*. The regression was also performed to include two features,* `logage` *and* `logassets`*, for each focal company. For each model, the first row reports on the results without these two features and the next row includes these two features. Note that the make-up of the peers does not change between these regression options.*

Table 2 drills down into the performance of the embedding over the website text for 3800 public companies (Panel 1), and the website text for an additional thirty-two thousand private companies (Panel 2). D2V+C500 shows the best performance in identifying In-Sector (IS) peers, and we explore this ground truth task further in the next section. We report R^2 and $RMSE$ results for profit prediction based on Net Operating Income over Total Assets (NOI/Assets or RNOA), labeled Profit / Assets, and the Net Operating Income over Sales, labeled Profit / Sales. We also assess the impact of two control variables: 1. logassets: the natural logarithm of the assets of the company (+1). 2. logage: the natural logarithm of the age of the company (+1), where age is measured as the number of years since the company's initial appearance in the Compustat database. Doc2Vec/Cosine, with an embedding over the 3800 public companies, has the highest R^2 values of 0.503 and 0.447 (Panel 1). D2V+C500 has somewhat lower values of 0.469 and 0.412 (Panel 2) The addition of the controls has a positive impact on predicting profitability, but does not impact the rank ordering of the various peer networks.

5.3. Financial Variables

We report on the ability of the D2V+C500 2%-threshold peers to predict / explain the following financial metrics commonly used in corporate finance and asset pricing research: 1. Market-to-Book (M/B) Ratio; 2. Market Leverage; 3. Book Leverage; 4. HML Beta; 5. SMB Beta.

Similar to the previous experiment, we report on a subset of 3800 public companies. We regress each dependent variable for the focal company (first column of Table 3) against the average value of the same variable for its D2V+C500 2%-threshold peers. The upper panel is a univariate regression, while the lower panel adds the assets and age of the focal company as controls. The left side of each panel reports on regressions that do not include self as a peer; this reflects the pure predictive power of the peers. The right side of each panel includes self as a peer and reflects the total level of explanatory power of the peers. The latter approach to include self resembles the norm for industry fixed effects models, where the focal firm is included in the respective industry group. We report on the peer effect coefficient (full transmission of the peer effect would result in a value of unity), the adjusted R^2, and $RMSE$. We cluster the standard errors at the company level. All of the coefficients are highly significant at the 99% confidence level.

The results indicate that website text based peers have a strong signal regarding their ability to predict / explain each of these financial metrics. The coefficients are near unity on the right side of the panel (include self as peer) indicating a full correlation without decay. We also find that the peer signal explanatory power does not change materially when adding controls for the assets and age of the focal company.

Table 3 Financial Variables - Explanatory Power

D2V+C500 over website text for 32K private companies						
Model	Without Self as a Peer			With Self as Peer		
	Coef	RSQ	RMSE	Coef	RSQ	RMSE
M/B Ratio	0.67	0.13	2.76	1.05	0.31	2.14
Market Leverage	0.77	0.25	0.03	1.01	0.44	0.02
Book Leverage	0.69	0.16	0.05	1.01	0.36	0.04
HML Beta	0.83	0.45	0.33	0.93	0.57	0.26
SMB Beta	0.53	0.07	0.48	1.01	0.28	0.37
D2V+C500 with logassets and logage over website text for 32K private companies						
Model	Without Self as a Peer			With Self as Peer		
	Coef	RSQ	RMSE	Coef	RSQ	RMSE
M/B Ratio	0.56	0.17	2.63	0.98	0.33	2.11
Market Leverage	0.72	0.28	0.03	0.97	0.45	0.02
Book Leverage	0.64	0.19	0.05	0.97	0.37	0.04
HML Beta	0.78	0.46	0.32	0.90	0.57	0.26
SMB Beta	0.49	0.08	0.47	1.00	0.28	0.37

The upper panel reports on baseline results using the peer network of the tuned WTNIC D2V+C500 and the lower panel reports on results that include the assets and age of the focal company as control variables. Both panels use the 2%-threshold for peers. The first column is the dependent variable. The next three columns on the left are the coefficient and the R^2 and RMSE values for regressions, where the dependent variable is regressed over the mean value for that variable, computed over the peers of each focal firm, but not including self as a peer. The next three columns on the right correspond to regressions where self is included as a peer.

5.4. In-Sector (IS) Peers

Table 4 provides insights into the ground truth task of (ii) identifying In-Sector (IS) peers in the competitor network. We define sectors using the broad industry grouping of Fama French 5 (Fama and French (1997)) as follows: (1) `(Consumer)` Durables and Non-Durables; Wholesale; Retail; Some Services. (2) `(Manufacturing)` Manufacturing; Energy; Utilities. (3) `(HiTec:)` Business Equipment; Telecommunications; TV; Control; Computational Services. (4) `(Health:)` Healthcare; Medical Equipment; Drugs. (5) `(Other:)` Multiple sub-sectors including Transportation and Construction; Finance. The `(Other)` group comprises 1303 or approximately one third of the 3800 companies in our ground truth data set. The other four groups range from 536 companies (14%) for `(Consumer)` to 678 companies (17%) for `(HiTec)`. We consider two thresholds to identify the closest peers; the first considers the Top 2% of all peers and the second considers the Top 1%.

Table 4 Distribution of IS Peers across Fama French 5 Sectors for Public Companies

Dataset	Model	All		ff5:1 (536)		ff5:2 (665)		ff5:3 (678)		ff5:4 (619)		ff5:5 (1303)	
		2%	1%	2%	1%	2%	1%	2%	1%	2%	1%	2%	1%
public	D2V	66	73	41	47	49	57	52	58	80	85	74	81
pub+32K	D2V	72	79	40	43	54	63	53	59	87	91	78	84
pub+32K	D2V+C500	92	97	73	79	80	90	67	82	97	99	94	98

This table displays the fraction of peers identified by each model (at 1% or 2% granularity) that are In-Sector (IS) peers. A higher ratio is good and indicates that the peers are well selected as peers in the same sector are more similar. Sectors 1 to 5 are Consumer, Manufacturing, Tech, Health, and Miscellaneous, respectively.

Table 4 reports the percentage of In-Sector (IS) peers for the three core web-based Doc2Vec approaches across the five sectors and we note clear differences. D2V (public and 32K private) outperforms D2V (public) across all five sectors in identifying IS peers. The latter has 66% IS peers (2%) and 73% IS peers (1%) overall, while the former has 72% (2%) and 79% (1%), respectively. This reflects a significant advantage of considering a much larger training corpus including the 32K private companies compared to just including the 3800 public companies.

Most crucial, we also see outstanding benefits to fine-tuning for D2V+C500 (3800 public and 32K private). This model has 92% IS peers (2%) and 97% IS peers (1%) overall. As we expected, the percentage of IS peers is highest for Sector (4), followed by Sector (5). With a 1% cutoff for D2V+C500 , the percentage of IS peers for Sector (4) `(Health)` is at an extraordinarily high level of 99%. The performance in identifying IS peers for companies in Sector (1) `(Consumer)` is also notable given this is challenging for the other models. This reflects the fact that firms in the consumer sector are more diverse than those in the health sector.

To summarize, D2V (3800 public and 32K private) outperforms D2V (3800 public), showing the advantage of training on the additional thirty two thousand private companies. We also see a significant improvement from the tuning of D2V+C500 (3800 public and 32K private), resulting in the identification of up to 99% of IS peers in one specification.

5.5. NAICS Prediction for Public and Private Companies

Tables 5 and 6 report on the performance of D2V+C500 to predict NAICS codes for focal public and private companies, respectively. Both tables display the fraction of peers that are in the same NAICS code as the focal company. We consider 2-digit to 6-digit NAICS granularities. The results are excellent and we emphasize that our work is the first to attempt such predictions for private companies.

The upper panel of Table 5 reports on the 2% network and the lower panel on the 1% network. D2V+C500 for the 1% network has 95.74% accuracy for a 2-digit match, which intuitively declines to 61.61% for a 6-digit match.

Table 5 NAICS Prediction for Public Companies

Model	Similarity	Network	Public count	2-digit	3-digit	4-digit	5-digit	6-digit
				Accuracy (%)				
Doc2Vec	Cosine	2%	3785	60.67	52.49	48.68	40.59	30.38
D2V+C500	Cosine	2%	3589	85.12	80.49	76.87	62.88	48.71
Doc2Vec	Cosine	1%	3664	69.94	63.38	60.02	50.15	38.55
D2V+C500	Cosine	1%	2095	95.74	94.46	92.87	72.67	61.01

This table displays the fraction of peers identified by each model that are in the same NAICS sector as the focal PUBLIC company. A higher ratio is good and indicates that the peers are well selected as peers in the same sector are more similar. We report on results as we vary the match from a 2-digit NAICS to a 6-digit NAICS code match.

For the private companies, Table 6 reports on a peer network of the Top 200K peers (upper panel) and the Top 100K peers (lower panel). We note that the Top 100K peer network shrinks the count of focal private companies, from 12074 (Top 200K) to 4980 (Top 100K); the count of the peer public companies also drops from 3582 (Top 200K) to 2087 (Top 100K). This shrinkage reflects that there is a wide range of quality of the website text for the private companies, and there are select companies that provide high quality website text that accurately reflects their products and services. Hence, the prediction accuracy is lower compared to that for public companies. D2V+C500 shows an accuracy greater than 50% for 2- and 3-digit peers for the Top 200K network, and a similar accuracy up to the 4-digit match, for the Top 100K network.

Table 6 NAICS Prediction for Private Companies

Model	Similarity	Network peer count	Private count	Public count	2-digit	3-digit	4-digit	5-digit	6-digit
					Accuracy (%)				
Doc2Vec	Cosine	200000	12074	3582	25.2	18.2	13.3	7.9	6.7
D2V+C500	Cosine	200000	11780	3171	64.69	54.30	36.42	22.3	19.9
Doc2Vec	Cosine	100000	9961	3164	31.70	24.86	19.50	11.32	9.9
D2V+C500	Cosine	100000	4980	2087	81.04	73.53	57.62	33.65	31.42

This table displays the fraction of peers identified by each model that are in the same NAICS sector as the focal PRIVATE company. A higher ratio is good and indicates that the peers are well selected as peers in the same sector are more similar. We report on results as we vary the match from a 2-digit NAICS to a 6-digit NAICS code match.

5.6. ETNIC Based Ground Truth

Going beyond ground truth based on SIC or NAICS codes, we consider a more stringent test. As noted earlier, ETNIC (Hoberg and Phillips (2025)) is a high-quality peer network that is based on an embedding constructed over the SEC 10-K filings. We compare the direct network overlap of Doc2Vec and D2V+C500 with ETNIC. The results are reported in Table 7, where we consider a threshold ranging from 2% to 20% of the networks. As can be seen, D2V+C500 significantly

outperforms Doc2Vec. The overlap with ETNIC is over 50% for the 2% threshold and is over 90% for the 20% threshold. These results are significantly stronger than those for un-tuned doc2vec, which range from 31% to 69%. Overall, these results are excellent since ETNIC is constructed over SEC 10-K filings, whereas D2V+C500 is constructed over the noisy website text. Moreover, they clearly illustrate the positive impact of fine-tuning to focus on website text describing products and services.

Table 7 Comparison Against ETNIC Based Ground Truth

Model	ETNIC Threshold (%)				IS 1%	IS 2%
	2	5	10	20		
Doc2Vec; Public	31.4	48.3	59	69	73	66
Doc2Vec; Public + 32K Private	33.6	51.0	62.8	68.7	77	72
D2V+C500; Public + 32K Private	51.5	73.4	83.9	90.9	97	95

This table displays the recall, i.e., the fraction of peers identified by each model that has an overlap with the ground truth of peers based on ETNIC competitor network (Hoberg and Phillips (2025)). We vary the threshold of the ETNIC network from 2% to 20%. A higher ratio is better.

5.7. LLM Ground Truth for Private and Public Companies

We next curate a human+LLM-generated ground truth of public to private peers as follows:

- We identified a random sample of 750 private companies that satisfied the following criteria: (1) They are available in business databases, e.g., Moody's Orbis or S&P Capital IQ, and were included in WTNIC; (2) They have SIC and/or NAICS code labels, indicating that they have been somewhat active.
- A human researcher visited the Website of each focal company and wrote a two sentence description of its products and services.
- We then queried the Claude LLM (specifically "claude-3-5-sonnet-20240620") with the following query:
  `"Each private company is identified by its Website URL and a two sentence brief business description of its primary product or service. Please return the Website URL name or ticker symbol of public companies listed on any exchange of public companies that sell the most similar product or service."`
- We then curated a ground truth dataset - ClaudeGT - to represent the responses from the Claude LLM. This database included approximately 7,500 entries, referencing 660 of the 750 private companies, and with 1,600 public companies as peers.
- We note that most generative AI models such as Claude lean toward choosing large(r) public companies that are more prevalent in their training datasets. For example, when we sort public companies into percentiles based on their assets, the 1,600 companies returned by Claude were on average assigned to the 75th percentile and higher. This limitation likely entails little if any bias in our tests. This test is also noteworthy because it does not rely on industry classifications such as NAICS that are known to have limitations.

Table 8 Claude Ground Truth for Private Companies

network threshold	Recall (fractional overlap)	
	Doc2Vec	D2V+C500
2%	30.1	37.1
5%	41	49.5
10%	51.1	58.1
20%	61.9	66.8

This table displays the recall, (in particular, the fractional overlap of recall), between the predicted peers for Doc2Vec and D2V+C500, and the human+LLM-generated ground truth peers obtained from Claude GPT. For all networks, D2V+C500 has significantly higher recall.

We create peer networks for the 660 private companies and the 1,600 public companies referenced in the ClaudeGT database, and we consider the Top 2% to the Top 20% subset of peers. Table 8 reports *recall* for the Top K% subsets for both the untuned Doc2Vec and fine-tuned D2V+C500. Note that *recall* measures the fractional overlap between the network's peers and the ground truth peers. We find that both Doc2Vec and D2V+C500 had relatively high values of recall ranging from 30% to almost 70%. Importantly, D2V+C500 outperformed Doc2Vec by a significant margin across all subsets of the peer networks.

6. Sector Wide Chinese Competition Shocks: Public vs Private Peers

The entry of low-cost products from China represents a large impact shock that affects entire industry sectors. This allows us to also explore the impact of these shocks separately, both to public and private companies. We note that little is known about how these shocks impact smaller private companies, particularly regarding how they reposition themselves in the product market. This exploration thus highlights a primary strength of the new WTNIC network, as it uniquely encompasses both public and private peers.

To construct the industry-level shock, we begin with a panel dataset of companies from 2000 to 2022. For these companies, we construct three variables, *Chinese Competition*, *High Chinese Competition* and *Chinese IP Competition*. These variables are computed using the MetaHeuristica platform, which contains the 10-Ks of public companies. For a focal public company and its 10-K filing, we consider the count of mentions of specific China shock keywords, scaled by the total word counts, following Hoberg et al. (2020). We then compute the average of these China shock variables over a focal company's TNIC peers to obtain a value for the exogenous industry-level China shock. This particular experiment requires a newly created full panel of network data from 2000 to 2022. [3] This is necessary because the impact of the China shock is identified using within-firm variation over time following the standards in the literature.

We then consider five dependent variables to assess the impact of these shocks on firms. The first two are the Generalized Hedonic-Linear Markup (GHL Markup) Pellegrino (2025) and the price markup over marginal cost proposed by De Loecker, Eeckhout, and Unger (DEU Markup) De Loecker et al. (2020). We also consider the following three dependent variables, ETNICpublic, D2Vpublic and D2Vprivate, representing the similarity score of the focal company to its peers. ETNICpublic is computed using the ETNIC peer network Hoberg and Philips (2024). D2Vpublic and D2Vprivate are computed using the new D2V+C500 network featured in this paper and are based on public company web peers and private company web peers, respectively. We consider the Top 2% of the network for TNICpublic and D2Vpublic, and the Top 1% of the network for D2Vprivate (a much larger network) as our baseline following earlier conventions in this paper. The dependent variable (total similarity score) is computed as the sum of the similarity scores with each of the peers of the focal company.

Table 9 displays the results and each coefficient reported is the result of a separate regression. All standard errors are clustered by company.

As expected, we first find that the China competition shock impacts markups. We also find that it results in higher total similarities with peers when similarity is measured using only public company peers. This is consistent with increased competition as reduced product differentiation reduces pricing power and also profits. But what is most novel is that the same China shock also results in **decreased** total similarity with the private company peers. This result is only possible to identify using WTNIC data, and it suggests that private companies may be more agile and are able to shift their products in ways that make them less impacted by the Chinese competitors and their influx of lower cost products. This result clearly illustrates that an important shock can lead to very different economic outcomes for public and private company peers, motivating wider use of WTNIC peer analysis for a wide array of questions in finance, economics, and beyond.

[3] The previous tables use only the 2016 network to show its relevance and applicability across multiple validation tests and methods.

Table 9 Regression Results for Industry Sector-Wide Chinese Competition Shocks

Dependent Variable	Chinese Competition	High Chinese Competition	Chinese IP Competition
GHL Markup	-0.063 (-8.51)	-0.096 (-8.64)	-0.134 (-5.64)
DLEU Markup	-0.039 (-1.98)	-0.063 (-2.01)	-0.037 (-0.56)
Private Company Similarity (D2V+C500)	-27.9 (-12.75)	-36.9 (-9.99)	-72.8 (-11.1)
Public Company Similarity (D2V+C500)	0.123 (6.43)	0.182 (6.47)	0.639 (8.33)
Public company Similarity (TNIC)	0.155 (11.04)	0.239 (10.74)	0.501 (10.63)

To examine the impact of this industry-level China shock on individual companies, we examine a panel dataset of public companies from 2000 to 2016. We consider three independent variables based on each public focal firm's industry peers, Chinese Competition, High Chinese Competition and Chinese IP Competition. These three RHS variables are computed from 10-Ks using the metaHeuristica platform as the average China shock of the focal firm's peers and thus represent more plausibly exogenous industry-level shocks regarding the rapid entry of low cost Chinese products. We consider five dependent variables. The first two are the Generalized Hedonic-Linear Markup (GHL Markup) and the price markup over marginal cost proposed by De Loecker, Eeckhout, and Unger (DEU Markup). We also consider three dependent variables, TNICpublic, D2Vpublic and D2Vprivate, representing the total summed similarity of the focal company to its peers (divided by 100 for ease of viewing).

7. Summary and Future Research

We present representation learning enabled approaches to generate the WTNIC peer competitor network. Our approach exploits non-proprietary public website text, and provides product and service focused coverage of both public and private firms within the same embedding space. We demonstrate the effectiveness of the WTNIC competitor network through an extensive evaluation of the quality of public and private peers. We document the ability of the identified WTNIC peers to explain important financial variables commonly used in corporate finance and asset pricing research. By placing the public and private companies within the same embedding space, WTNIC will facilitate research on the evolution of both public and private companies, as well as the influence of private competitors on public firms.

We make the following contributions based on our extensive evaluation of the WTNIC network over a range of datasets, and approaches to construct and tune an embedding:

- Our proposed best solution which results in the WTNIC network is built on a Doc2Vec embedding (Mikolov et al. (2013)) over website text, that is further fine-tuned to ensure the peer network focuses on website content specific to products and services. We evaluate the WTNIC network using multiple ground truth datasets since identifying the most accurate product market competitors is our core objective. We find that WTNIC competitors outperform the existing TNIC benchmark, especially on our wide array of competitor ground truth tests. WTNIC provides a robust platform for constructing the first unified competitor network for both public and private companies.
- High quality peer networks should rarely identify product market competitors (peers) that are not in the same broadly defined industry sector. The use of the coarse Fama French 5 industry sectors makes failing this ground truth test stark. Our proposed fine-tuned WTNIC D2V+C500 model provided the best performance for this ground truth test. The WTNIC network also performed remarkably well at identifying peers based on NAICS codes based ground truth, at varying levels of granularity. WTNIC is the first network to predict public peers for private companies. These results also reflect the benefits of the WTNIC training on the website text of both public and private companies instead of public companies alone.
- We introduce a novel and critical test based on a set of *human+LLM-generated ground truth peers*. For each focal private company in a random sample, we ask a Claude LLM to identify those publicly traded companies that are its most similar product market peers. We compute the recall (fraction of overlap) between the predicted peers and the human+LLM-generated peers. Our baseline fine-tuned WTNIC network performs very well, and provided the best performance on this task.

Contributions from a financial/economic perspective are demonstrated in two ways. The WTNIC network can predict financial variables commonly used in corporate finance and asset pricing research. This includes the Market-to-Book (M/B) Ratio, Market Leverage, Book Leverage, Market

Beta, HML Beta, and SMB Beta. The results suggest that WTNIC has significant ability to explain each of these variables. Its explanatory power does not change materially when adding controls for size and age, indicating that the WTNIC signal is important and stable.

The WTNIC contribution is also shown using an experiment assessing the impact of Chinese competition shocks on both the public and private peers of a focal company. We find that private and public firm responses are quite different and are consistent with improved private firm agility. These results are novel, as little is known about how these shocks affect private companies and their spatial positioning in the product market. This exploration underscores the primary strength of the WTNIC network, which includes both public and private peers.

WTNIC also makes contributions regarding the methodology to construct and fine tune an embedding. We find that a fine tuned Doc2Vec performs well and has good properties regarding robustness and efficiency in this setting. A key challenge is that website text is noisy. Token sequences may not resemble sentences and may not be meaningful. This is unlike other corpora such as SEC 10-K filings. This issue is relevant because the transformer models that we examine, including RoBERTa (Devlin et al. (2019)), LongFormer (Beltagy et al. (2020)) and FinBERT (Araci (2019)), are typically trained on sentences. We note that the robustness and efficiency of Doc2Vec depends on the ability to easily train it on website text, and that it is scalable to our very large corpus with several hundred thousand private company websites.

Finally, the WTNIC network is free of look-ahead bias. We recreate the network separately for each year, only using information that is available as company website text within the past three years. We view this latter issue of look-ahead bias to carry high weight in our methodological choices. In particular, we aim to be conservative and provide guarantees about reducing look-ahead bias, thereby facilitating a broader set of future use cases for the WTNIC network.

We briefly discuss future research. Potential challenges include alignment across diverse embedding spaces and the construction of multiple distinct competitor networks. Website text includes content beyond products and services, such as employment information, corporate governance, and Enviromental, Social and Governance (ESG) information. We note that website text is only one of many sources of textual data. Companies increasingly use social media platforms to communicate, and future researchers might further supplement website text with content from social media platforms such as LinkedIn, Facebook, and Yelp.

Given this rich content, one could construct additional specialized embedding spaces with a focus on (i) patents, (ii) ESG and corporate governance, (iii) job descriptions, etc. Alignment across these embedding spaces will enable insights, e.g., one could learn about relationships, such as the concentration of specific job categories within certain industry sectors. A second and related challenge would use these embeddings and labeled data to construct multiple distinct competitor networks. For example, the peers of a focal company in the product space may differ substantially from its peers in the labor market or from its competitors when seeking patents.

Appendix

Reproducible Research: This paper describes the construction and evaluation of a website text based network industry classification (WTNIC) database, spanning the years 2000 to 2025, and including approximately 500,000 public and private companies. The WTNIC database, and the fully documented scripts to create the database, will be made available under an appropriate license for free download.

Acknowledgements: We thank the National Science Foundation (grants 1561068 and 1937153), the Dartmouth Tuck School of Business, the USC Marshall Institute for Outlier Research in Business (iORB) and the Robert H. Smith School of Business at the University of Maryland for supporting this research.

References

Acikalin U, Caskurlu T, Hoberg G, Phillips G (2022) Intellectual property protection lost and competition: An examination using large language models. Tuck School of Business Working Paper No. 4023622 URL http://dx.doi.org/10.2139/ssrn.4023622.

Araci D (2019) Finbert: Financial sentiment analysis with pre-trained language models. URL http://dx.doi.org/10.48550/arXiv.1908.10063.

Bao S, Li R, Yu Y, Cao Y (2008) Competitor mining with the web. IEEE Transactions on Knowledge and Data Engineering 20(10):1297–1310, URL http://dx.doi.org/10.1109/TKDE.2008.98.

Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer. arXiv:2004.05150 .

Bernstein A, Clearwater S, Hill S, Perlich C, Provost F (2002) Discovering knowledge from relational data extracted from business news. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on MultiRelational Data Mining.

Bhojraj S, Lee CMC (2002) Who is my peer? a valuation-based approach to the selection of comparable firms. Journal of Accounting Research 40(2):407–439.

Bhojraj S, Lee CMC, Oler DK (2003) What's my line? a comparison of industry classification schemes for capital market research. Journal of Accounting Research 41(5):745–774.

Blei D, McAuliffe J (2007) Supervised topic models. Proceedings of the 20th International Conference on Neural Information Processing Systems, 121–128, NIPS'07, ISBN 9781605603520.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993–1022.

Chamberlin E (1933) A Theory of Monopolistic Competition (Harvard University Press).

Cong W, Liang T, Zhang X (2024) Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. Cornell University Working Paper .

De Loecker J, Eeckhout J, Unger G (2020) The Rise of Market Power and the Macroeconomic Implications*. The Quarterly Journal of Economics 135(2):561–644, URL http://dx.doi.org/10.1093/qje/qjz041.

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota: Association for Computational Linguistics), URL http://dx.doi.org/10.18653/v1/N19-1423.

Fama EF, French KR (1997) Industry costs of equity. Journal of financial economics 43(2):153–193.

Hanley K, Hoberg G (2019) Dynamic interpretation of emerging risks in the financial sector. Review of Financial Studies URL http://dx.doi.org/10.1093/ssrn.2792943.

Hilt V, Schwenkler G (2020) The different networks of firms implied by the news. URL http://dx.doi.org/10.2139/ssrn.4946066.

Hoberg G, Li Y, Phillips G (2020) Internet access and u.s. - china innovation competition. URL http://dx.doi.org/10.3386/w28231.

Hoberg G, Philips GM (2024) Scale, Scope and Competition: The 21st Century Firm. Journal of Finance Forthcoming.

Hoberg G, Phillips G (2025) Scope, scale and concentration: The 21st century firm. Journal of Finance URL http://dx.doi.org/10.2139/ssrn.3746660.

Hoberg G, Phillips GM (2010) Product market synergies and competition in mergers and acquisitions: A text-based analysis. Review of Financial Studies URL https://doi.org/10.1093/rfs/hhq053.

Hoberg G, Phillips GM (2016) Text-Based Network Industries and Endogenous Product Differentiation. Journal of Political Economy 124(5).

Hofmann T (2017) Probabilistic latent semantic indexing. ACM SIGIR Forum 51(2):211–218.

Hotelling H (1929) Stability in competition. Economic Journal 39(153):41–57.

Jha M, Qian J, Weber M, Yang B (2024) Chatgpt and corporate policies. Georgia State University Working Paper .

Kolbel J, Leippold M, Rillaerts J, Wang Q (2024) Ask BERT: How Regulatory Disclosure of Transition and Physical Climate Risks affects the CDS Term Structure. Journal of Financial Econometrics 22(1):30–69.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. International Conference onmMachine Learning, 1188–1196.

Lee C, Ma P, Wang C (2015) Search-based peer firms: Aggregating investor erceptions through internet co-searches. Journal of Financial Economics URL http://dx.doi.org/10.1016/ssrn.2171497.

Lee J, Stevens N, Han SC, Song M (2024) A survey of large language models in finance (finllms). URL http://dx.doi.org/10.48550/arXiv.2402.02315.

Li Y, Wu X (2025) Labor links, comovement, and predictable returns. Journal of Financial and Quantitative Analysis URL http://dx.doi.org/10.1017/ssrn.3175958.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach.

Liu Z, Huang D, Huang K, Li Z, Zhao J (2021) Finbert: a pre-trained financial language representation model for financial text mining. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), ISBN 9780999241165.

Loper E, Bird S (2002) Nltk: The natural language toolkit. In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

Lopez Lira A, Tang Y (2024) Can chatgpt forecast stock price movements? return predictability and large language models. University of Florida Working Paper .

Ma Z, Pant G, Sheng OR (2011) Mining competitor relationships from online news: A network-based approach. Electronic Commerce Research and Applications 10(4):418–427.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 3111–3119.

Office of Management and Budget (2017) North American Industry Classification System (United States Census).

Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, Grangier D, Auli M (2019) fairseq: A fast, extensible toolkit for sequence modeling. Proceedings of NAACL-HLT 2019: Demonstrations.

Pant G, Sheng O (2015) Web footprints of firms: Using online isomorphism for competitor identification. Information Systems Research 26(1):188–209, URL http://dx.doi.org/10.1287/isre.2014.0563.

Pearce E (1957) History of the Standard Industrial Classification (Bureau of the Budget, Office of Statistical Standards).

Pellegrino B (2025) Product Differentiation and Oligopoly: A Network Approach. American Economic Review 115:117–1225, URL http://dx.doi.org/10.1257/aer.20201425.

Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543.

Rauh JD, Sufi A (2011) Explaining corporate capital structure: Product markets, leases, and asset similarity. Review of Finance 16(1):115–155.

Rehurek R, Sojka P (2011) Gensim–python framework for vector space modelling. Technical Report, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3(2).

Roberts M, Stewart B, Tingley D, Airoldi E (2013) The structural topic model and applied social science. International Conference on Neural Information Processing.

Wei Q, Qiao D, Zhang J, Chen G, Guo X (2016) A novel bipartite graph based competitiveness degree analysis from query logs. ACM Trans. Knowl. Discov. Data 11(2), ISSN 1556-4681, URL http://dx.doi.org/10.1145/2996196.

Xu K, Liao SS, Li J, Song Y (2011) Mining comparative opinions from customer reviews for competitive intelligence. Decision Support Systems 50(4):743–754, ISSN 0167-9236, URL http://dx.doi.org/https://doi.org/10.1016/j.dss.2010.08.021.

Yang Y, UY MCS, Huang A (2020) Finbert: A pretrained language model for financial communications. URL http://dx.doi.org/10.48550/arXiv.2006.08097.

Yaros JR, Imieliński T (2015) Data-driven methods for equity similarity prediction. Quantitative Finance 15(10):1657–1681, URL http://dx.doi.org/10.1080/14697688.2015.1071079.