Text-Based Network Industries and Endogenous Product Differentiation

Gerard Hoberg, Gordon Phillips

Reporter: Haimo Fang
School of Economics
Fudan University

July 18, 2025

Before We Begin...

High level summary: [Hoberg and Phillips, 2016] used text from each firm's 10-K product description to build a dynamic, network-based map of who really competes with whom.

Road-map

Motivation

Data & Pipeline

Methodology

Validation

Applications

Takeaways & Limits

Why Defining Industries Matters

- *Investors & Analysts:* Accurate peer sets improve relative valuation multiples, risk benchmarking, and portfolio diversification.
- *Managers:* Clearer view of the competitive landscape guides strategic positioning, pricing, and R&D focus.
- *Job Seekers & Workers:* Understanding industry boundaries informs career planning and mobility prospects.
- *Antitrust & Merger Review:* Market-definition tests hinge on identifying the set of close competitors.
- *Industrial Policy:* Targeted subsidies, tariffs, or tax incentives require precise sector delineation.
- *Economic Measurement:* GDP, productivity, and inflation statistics rely on coherent industry classifications.

Definition of Industry(Conventional Code)

- **SIC (Standard Industrial Classification)**
  - Introduced in the 1930s by the U.S. government
  - Four-digit, production-based hierarchical codes
  - Rarely updated; enforces rigid, transitive groupings
- **NAICS (North American Industry Classification System)**
  - Adopted in 1997 by U.S., Canada, and Mexico
  - Six-digit codes aligned with modern production processes
  - Revised quinquennially; remains static and coarse
- Both treat industries as fixed, non-overlapping "boxes" that can be slow to adapt to new products or shifting competition.

Definition of Industry (cont.)

- *Primary activity basis:* each firm self-reports its main revenue source in surveys/filings
- **SIC:** 4-digit code
  - 2 digits = major industry group (e.g. "20" = Food & Kindred Products)
  - 3rd digit = industry subgroup (e.g. "201" = Meat Products)
  - 4th digit = specific industry (e.g. "2011" = Meat Packing)
- **NAICS:** 6-digit code
  - 2 digits = economic sector (e.g. "31" = Manufacturing)
  - 3 digits = subsector (e.g. "311" = Food Manufacturing)
  - 4 digits = industry group (e.g. "3115" = Dairy Product Mfg)
  - 5 digits = NAICS industry (e.g. "31151" = Dairy Product Mfg)
  - 6 digits = national detail (e.g. "311511" = Fluid Milk Mfg)

Limitations of SIC & NAICS

Slow to update

- SIC dates to the 1930s; NAICS only refreshes every five years
- New industries (e-commerce, biotech spin-outs) lag behind

Rigid, transitive groupings

- A–B and B–C linkage forces A–C, even if they're unrelated

Coarse buckets only

- Wildly different firms can share the same 4- or 6-digit code
- No continuous "how close?" measure

Production-process focus

- Groups by inputs/activities, not by products or customer markets

Objective

SIC/NAICS: Define industries based on production processes
10-K fixed cluster / TNIC: Place firms in defined industry based on the product they offered to the customers.

- Provide a new measure of similarity without detailed product prices and quantities (IO research conventional approach to compare cross-price elasticities)
- Allow frequent annual updating of industry definition
- Capture horizontal relatedness between firms, not vertically

Data & Text Pipeline: Overview

1. **Collect 10-K Business Descriptions**
2. **Clean & Tokenize Text**
3. **Vectorize & Compute Similarity**
4. **Build Clusters & Networks**

(We'll walk through each step in detail.)

Corpus: 10-K Business Descriptions

- **Universe:** All U.S. public firms with annual 10-K filings (1996–2008), covering 98 % of CRSP/Compustat firms
- **Source sections:** Item 1 ("Business") — scraped from EDGAR (10-K, 10-K405, 10-KSB, 10-KSB40)
- **Documents:** 50,673 firm-year business descriptions; firms with <20 unique words excluded; typical firm uses 200 unique words (range: 50–1,000)
- **Word selection:**
  - Nouns (Webster.com) & proper nouns (capitalized 90 % of uses)
  - Drop words appearing in >25 % of firms, and geographic terms (all country/state names, top 50 cities)
- **Vocabulary size:** $W = 61{,}146$ unique nouns/proper nouns in 1996, $55{,}605$ in 2008; each firm-year $\rightarrow$ binary W-vector

Distribution of Word Frequencies of Firms

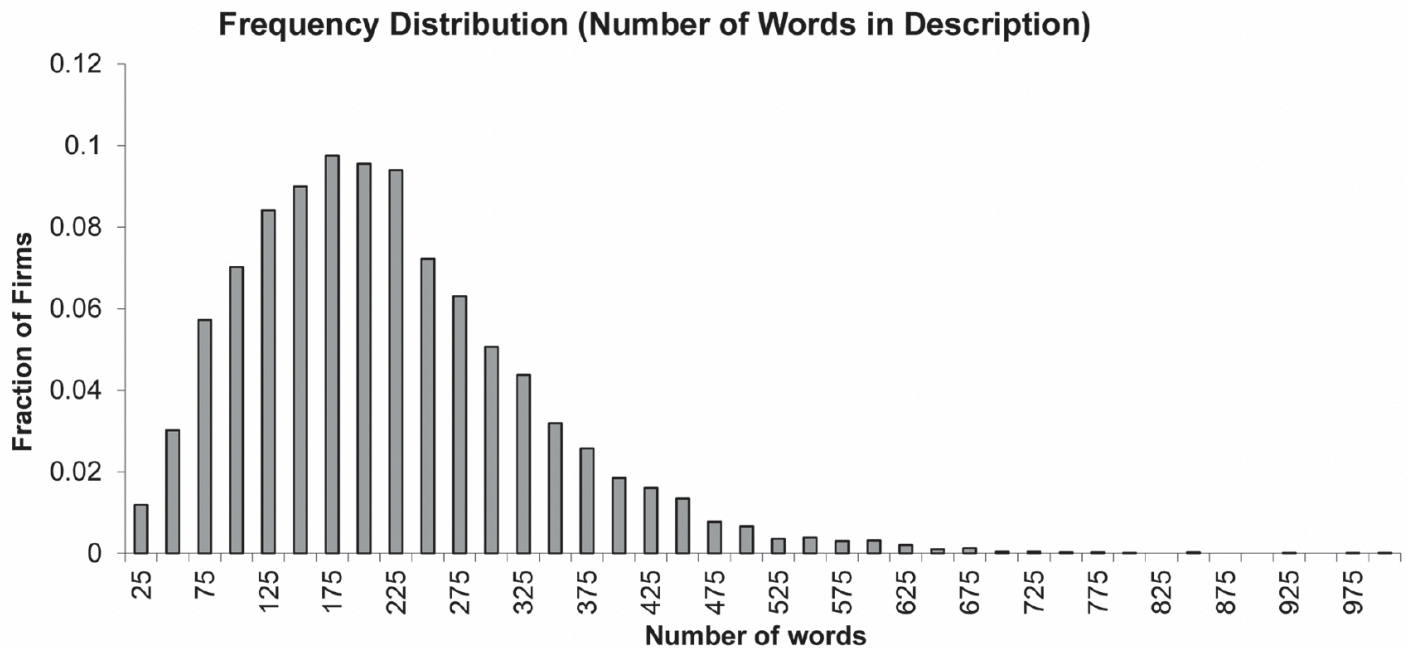

FIG. 1.—Frequency distribution of unique non–common noun and proper noun words in 10-K product descriptions. Color version available as an online enhancement.

Text Cleaning & Tokenization

- **POS filtering**
  - Keep only nouns (matched against a standard dictionary)
  - Keep proper nouns: words capitalized in 90 % of their occurrences
- **Boilerplate removal**
  - Discard any term that appears in more than 25 % of all firm-year filings (e.g. "Company," "Business," "Segment")
- **Stop-word & geographic filter**
  - Remove standard stop-words (e.g. "the," "and," "of")
  - Exclude all country/state names and the top 50 city names
- **Tokenization & normalization**
  - Split text on whitespace and punctuation
  - Convert to lowercase
  - Drop tokens of length < 2 characters
- **Vocabulary construction**
  - No stemming or lemmatization—treat each word form distinctly
  - Result: a high-dimensional binary vector indicating presence/absence of each cleaned token

Vectorization & Cosine Similarity

- **Binary presence vector:**

$$P_i \in \{0,1\}^W, \quad W \approx 60{,}000$$

indicates which cleaned tokens appear in firm i's description.

- **Normalization:**

$$V_i = \frac{P_i}{\|P_i\|_2}, \quad \|V_i\|_2 = 1$$

- **Cosine similarity:**

$$s_{ij} = V_i^\top V_j = \frac{|P_i \cap P_j|}{\sqrt{|P_i|\ |P_j|}} \in [0,1]$$

- **Result:** A full $N_t \times N_t$ similarity matrix M_t each year, which becomes the input for clustering (10K-Fixed) and network-thresholding (TNIC).

Vectorization & Cosine Similarity (cont.)

10K-Fixed Clustering (base year 1997)

- **Input:** single–segment firms' unit vectors $\mathbf{V}_i \in \mathbb{R}^W$ and similarity matrix $S = (s_{ij})$ with $s_{ij} = \mathbf{V}_i^\top \mathbf{V}_j$.
- **Average-linkage agglomeration:**

$$\bar{s}(A, B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} s_{ij};$$

$$(A_t, B_t) = \arg \max_{A \neq B, \{A,B\} \in \binom{\mathcal{C}^{(t)}}{2}} \bar{s}(A, B).$$

  Merge the most similar pair (A_t, B_t) until $K = 300$ clusters remain.
- **Single-swap refinement:** move any firm if it increases total within-cluster similarity $W(\mathcal{C}) = \sum_C \sum_{i<j \in C} s_{ij}$.
- **Stationary assignment:** for $t > 1997$ allocate firm i to $\arg\max_C \mathbf{V}_i^\top \bar{\mathbf{V}}_C$ where $\bar{\mathbf{V}}_C$ is the 1997 centroid.
- **Outcome:** ≈ 300 mutually exclusive, transitive industries fixed across all years.

10K-Fixed Clustering (cont.)

Text-based Network Industry Classification (TNIC)

- **Input each year t:** similarity matrix $S_t = (s_{ij})$. Centre row-wise to reduce length bias (generic words $\rightarrow$ higher raw similarity)

$$m_i = \text{median}_{j \neq i}\, s_{ij}, \qquad z_{ij} = s_{ij} - m_i.$$

- **Global threshold:** $\theta = 0.2132$ chosen so the fraction of *membership pairs* equals 2.05 % (3-digit SIC density).
- **Directed adjacency:** $a_{ij}(t) = \mathbf{1}\{z_{ij} \geq \theta\}$.
- **Firm-centric industry:**

$$\mathcal{I}_i(t) = \{\, j \neq i : a_{ij}(t) = 1\} \cup \{i\}.$$

  Industries are asymmetric and need not be transitive.
- **Dynamic:** similarities, medians, and peer sets are recomputed every year, tracking shifting competitive landscapes.

TNIC (cont.)

Threshold Selection (SIC Benchmark)

- **Compute SIC-3 density:** $d_{\text{SIC3}} = \frac{\#\{i<j\text{:3-digit SIC}_i=\text{SIC}_j\}}{\binom{N}{2}}$
- **Quantile rule:** Choose τ_t as the $(1 - d_{\text{SIC3}})$-quantile of the distribution of all s_{ij} in year t.
- **Why this matters:** Ensures 10K-Fixed clusters and TNIC peer-sets have the same "aggressiveness" (average rival count) as SIC-3.
- **Robustness:** The paper checks thresholds from the 90th to 99th percentile—results hold across this range.

Validation: Overview

- **Cross-industry dispersion** – Does text-based grouping produce tighter, more economically meaningful buckets?
- **MD&A "competition" mentions** – Do managers in high-similarity peer-sets write about intense competition more often?
- **Capital IQ peer-set match** – How well do TNIC peers align with self-declared rivals in an external database?

Cross-Industry Dispersion

- Compute standard deviation of operating income-to-sales (OI/Sales) across groups:
  - NAICS-4: 0.205
  - 10K-Fixed: 0.231
  - TNIC (unweighted): 0.248
  - TNIC (similarity-weighted): 0.267
- Similar increases found for sales growth and stock-return beta dispersion
- *Interpretation:* Text-based industries carve firms into more homogeneous clusters

MD&A "Competition" Mentions

- Regress indicator of mention on average peer similarity:

$$\Pr(\text{“competition”}) = \alpha + \beta\, \overline{s}_{i\cdot} + \varepsilon$$

- Key result: $\beta > 0$ and highly significant – Firms with tighter TNIC peer-sets are more likely to report "intense competition"
- Effect persists even when peers are in the 2–5 % similarity band (latent threats)

Match to Capital IQ Peer-Sets

- Compare TNIC-defined peers to self-declared rivals in Capital IQ
- Match rates (fraction of CapIQ rivals recovered):
  - SIC/NAICS: 44–47
  - TNIC: 52–55
- *Takeaway:* TNIC not only yields tighter clusters but also better aligns with managers' own view of competition

Applications: Exogenous Shocks

- **9/11 Defence Demand Shock**
  - Post-September 2001, average TNIC similarity among defence & aerospace firms jumps markedly
  - Number of TNIC peers and network density both rise sharply
  - Vocabulary shifts toward military-related terms (e.g. "defense," "intelligence," "logistics")
- **Dot-Com Bust (2000–2001)**
  - Internet/software firms experience a pronounced drop in similarity
  - Rival sets scatter into specialized niches (e-mail, Linux, security)
  - Illustrates how competitive groups reconfigure dynamically in response to market shocks

Applications: Endogenous Product Differentiation

- **Differentiation via Sunk Costs**

$$\Delta s_{i,t+1} = \alpha + \beta_1 \frac{\text{R\&D}_i}{\text{Sales}_i} + \beta_2 \frac{\text{Adv}_i}{\text{Sales}_i} + \varepsilon$$

- $\beta_1,\ \beta_2 < 0$: firms with higher R&D or advertising intensity become less similar to peers next year
- Consistent with Sutton's theory: sunk-cost investments create differentiation and boost profit margins

Applications: Within-Industry Heterogeneity

- **Uncovering Submarkets in Broad Codes**
  - SIC 737 "Business Services" hides six distinct TNIC clusters
  - Clusters include software, online retail, medical IT, consulting, HR services, document management
  - Cross-cluster similarity is low, revealing hidden specialization
- Shows how TNIC can dissect broad industry buckets into economically meaningful niches

Key Takeaways

- **Dynamic, data-driven industries:** Replaces static SIC/NAICS boxes with year-by-year text-based groupings.
- **Continuous similarity measure:** Cosine on binary word vectors yields a "how close?" metric.
- **Two classification flavors:**
  – 10K-Fixed: modern analogue of SIC-3 clusters
  – TNIC: overlapping, ego-network peer sets that capture nuanced competition.

Key Takeaways

- **Broad applicability:**
  – Better cross-industry dispersion of performance metrics
  – Aligns more closely with managers' own stated rivals
  – Captures dynamic responses to shocks and strategic differentiation
- **Open data resource:** TNIC data publicly released and widely adopted in IO, finance, and strategy research.

Limitations & Caveats

- **Public firms only:** Omits private and emerging competitors—may understate true rivalry.
- **Text manipulation risk:** Firms could game 10-K language, so not yet suitable for regulatory use.
- **Binary vector choice:** Ignores term frequency and semantics beyond noun presence; TF–IDF tested but under-performed.
- **Threshold calibration:** Relies on SIC-3 density as benchmark—assumes SIC peer counts are optimal.
- **Future directions:**
  - Incorporate more advanced NLP (embeddings, topic models)
  - Extend to private firms or international filings
  - Explore dynamic thresholds that adapt to market volatility

References


Hoberg, G. and Phillips, G. (2016).
Text-based network industries and endogenous product differentiation.
Journal of Finance, 71:123–170.

Thank You

Thank you!

Questions?