Text-Based Representations of Market Structures

Gerard Hoberg
University of Southern California
Marshall School of Business

Swissquote Conference [Nov 2018]
Concept of “industry” is undeniably important

* Which firms are industry rivals is most core to market structure.
* But world far more complex than “all or nothing” categorizations.
* What about close-rivals, partial-rivals, entry-threat rivals or complementary firms.
Market structure much deeper:

- Also: Product differentiation.
- Also: Innovation and Change.
- Also: Supply chains.
- Also: Economies of scope & entry.

All are seminal to how markets work and evolve. Goal is to design a unified empirical framework incorporating all of the above.
Section I: Market Structure PAST

Section II: Market Structure PRESENT

Section III: Impact and Uses

Section IV: Market Structure FUTURE

Section V: Using Text to Model Innovation
Govt employees gathering industry data for SIC codes. This is how things were 50 years ago!
Researchers and practitioners typically identify industries using govt classifications such as SIC, NAICS. These are constrained to be “transitive” & “binary”.

![SIC code industries network](image)
Limitation 1: Difficult to model change

* How can a company model the aftermath of disruption?
* Can a firm detect a rival’s gradual encroachment?
* Only a 100% transformation can justify reclassification (RARE).
Limitation 2: Cannot model product heterogeneity

*SIC models all products in an industry as 100% identical.  
*iPhones and Androids are exactly the same, right?
Limitation 3: Cannot model economies of scope

Can firms identify markets with synergies or “low cost” entry.

Cannot model entry threats either. No information in SIC codes!

* Can firms identify markets with synergies or “low cost” entry.
* Cannot model entry threats either. No information in SIC codes!
Section I: Market Structure PAST

Section II: Market Structure PRESENT

Section III: Impact and Uses

Section IV: Market Structure FUTURE

Section V: Using Text to Model Innovation
Computers are the most loyal work force ever. They only want electricity! Use scaleable NLP to create generalized concept of industry.
Fresard, Hoberg and Phillips (2016WP): Vertical relatedness in TNIC space and links to mergers and innovation.
See website www.marshall.usc.edu/industrydata (linked from my homepage at USC).
Parse 10-K Filings from the SEC

Extract Item 1 (Business Description) from every 10-K in each year.
Use Cosine Similarities

Document Similarity

- Doc 1: “We sell digital music products and computers.”
- Doc 2: “We sell smart phones that use digital data to consumers.”

- Step 1) Drop common words “we”, “sell”, “to”, “for”, “their”, “products” (identified globally).
- Step 2) 5 elements: “computers”, “consumers”, “digital”, “music”, “phones”

\[
P_1 = (1,0,1,1,0) \quad P_2 = (0,1,1,0,1)
\]

- Step 3) Normalize vector to have unit length of 1:

\[
V_1 = (0.577,0.577,0.577,0) \quad V_2 = (0,0.577,0.577,0.577)
\]

- Step 4) Compute cosine similarity \( V_1 \cdot V_2 = 0.3333 \)

- This dot product has a natural interpretation: \( \cos(\theta) = \)

- Cosine similarity is bounded between (0,1)

Key: map firm text to vectors of length one in Euclidean space! Firms have product market “locations” on a unit sphere.
Product Market Space is a high dimensional Unit Sphere

Every firm has a location on this sphere. So it has 5000+ public firms and is 80,000 dimensional. Firms residing in dense areas face high levels of competition. Firms in isolated parts of the space are differentiated and effective monopolies.
Network nearly unconstrained: it is intransitive and non-binary. All values in \([0, 1]\) (partial product overlaps) are possible.
Consequence 1: Can model changing industries!

Network fully redrawn each year. Small and large changes observable. We find massive annual change. 30% of peers this year are not next year! SIC-code peers are nearly 100% persistent.
Consequence 2: **Can** model product diff. & entry threats

Complete information about product differentiation. Positive but lower similarities indicate scope for entry & synergies.
Validation is Strong. Superior signal on profitability and stock returns

TNIC Industry Classifications
Adjusted RSQ: various characteristics

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>SIC3 Fixed Eff.</th>
<th>SIC3 Ind.-Yr Avg</th>
<th>NAICS 4 Fixed Eff.</th>
<th>10-K based Fixed Eff.</th>
<th>10-K FIC Ind-Yr Avg</th>
<th>10-K based TNIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating Income/Sales</td>
<td>28.4%</td>
<td>31.2%</td>
<td>28.7%</td>
<td>32.7%</td>
<td>37.2%</td>
<td>45.8%</td>
</tr>
<tr>
<td>Adver./Sales</td>
<td>4.1%</td>
<td>8.4%</td>
<td>6.1%</td>
<td>7.1%</td>
<td>16.9%</td>
<td>27.2%</td>
</tr>
<tr>
<td>Market Beta</td>
<td>9.6%</td>
<td>15.3%</td>
<td>9.7%</td>
<td>10.4%</td>
<td>15.7%</td>
<td>24.5%</td>
</tr>
</tbody>
</table>

Conclude: Reduces error in variables problem with industry controls
1) Benefits most substantial when complete intransitive network is used

Not only is TNIC more general in modeling of market structure. It is 50% more informative than SIC! Validation is critical.
*SIC-code momentum weak out after Moskowitz and Grinblatt (1999).
*Even basic TNIC momentum strategies reliably profitable.
*See Hoberg and Phillips JFQA (forthcoming).
Section I: Market Structure PAST

Section II: Market Structure PRESENT

Section III: Impact and Uses

Section IV: Market Structure FUTURE

Section V: Using Text to Model Innovation
Welcome to the Hoberg-Phillips Data Library

<< NEW: All TNIC and Fluidity data now extended through 2016! >>

<< Also, more granular TNIC data is also available >>

Data provided by Gerard Hoberg (USC) and Gordon Phillips (Dartmouth)

Text-based Network Industry Classifications (TNIC) data [click here]

* These new industry classifications are based on firm pairwise similarity scores from text analysis of firm 10K product descriptions. Competitors are firm-centric with their own distinct set of competitors - analogous to a network of a Facebook circle of friends. These new industry classifications are updated annually and offer research flexibility and are also more informative than FIC (fixed industry) classifications such as SIC, NAICS, and the 10-K based FIC classifications below. Our research shows they sharpenly improve upon SIC and NAICS codes in explaining many different firm-specific decisions, including firm profitability, Tobin's Q and dividends. The classifications are outlined in Hoberg and Phillips (2010, 2018), with references available by clicking on above link.

Industry Concentration and Total Similarity Data [click here]

* HHI Concentration metrics and Total Similarity data is available based on TNIC industries.

Product Market Fluidity Data [click here]

* Product Market Fluidity data assesses the degree of competitive threat and product market change surrounding a firm, and is based on Hoberg, Phillips and Prabhu (2014).

Now at: www.marshall.usc.edu/industrydata
Monthly volume steadily increasing

Now roughly 500 visitors per month. 22,000 total since inception.
Users are global!

*Switzerland is ranked 99th in population but ranks 10th on TNIC website visits! That beats Japan, Italy, and India!
Visitors hail from all 50 states

Yes, even North Dakota!
Who are these users?

* Not just for academic research!
* Industry users mostly I-banks, hedge funds, & tech firms.
Section I: Market Structure PAST

Section II: Market Structure PRESENT

Section III: Impact and Uses

Section IV: Market Structure FUTURE

Section V: Using Text to Model Innovation
Semantic relatedness can improve power beyond cosine similarities.
Embedding technologies showing strong promise in pilot study.
Latent Dirichlet Allocation not recommended for economic reasons.

* New improved TNIC models coming soon. Stay tuned!
Major Expansion to Private Firms

Current Researchers

WTNIC
Web Text-based Network Industry Classifications

Center on Knowledge Graphs
Information Sciences Institute

NSF

We thank the National Science Foundation!
$500,000 Grant
Joint B-School & USC’s Viterbi School (ISI)

Joint NSF-funded research b/t B-School and Comp Sci Researchers.
Get Annual Snapshots of 800,000 Private Firm Websites

(1) Gather and process web pages, (2) extract product market text, (3) store on cloud, and (4) code up scaleable NLP to compute network.
Very large network

* 18 years and 800,000 firms. Time varying and intransitive non-binary network.
* Use of standard similarity methods not feasible.
* Use scaleable sparse technologies and dimensionality reductions (embeddings).
Private firms are just very numerous

Only 1% are public firms

The rest of the network is a vast ocean of essentially unstudied private firms.

Many are VC-backed. Others are stable but private. Goal is to explore network evolution especially of early stage firms

Standard TNIC will be dwarfed by WTNIC. Understanding private firm entry and evolution invaluable to corporate managers, investors, and researchers.
Excellent Progress on WTNIC Execution

WTNIC network will be dynamic and redrawn each year.

Each year will have ~300,000 firms. That means 90 Billion pairs!

In contrast, TNIC public firm network has ~6,000 firms per year, which is only 36 million pairs.

WTNIC, needs scaleable tech from cutting edge comp sci researchers (USC’s ISI group!)

Alpha version of network only 3-6 months away!
Section I: Market Structure PAST

Section II: Market Structure PRESENT

Section III: Impact and Uses

Section IV: Market Structure FUTURE

Section V: Using Text to Model Innovation
Innovation: Patent Markets

Bowen, Fresard and Hoberg (2018)

- > 8 million patents since 1900. Text structure: abstract – description – claim

*Firms compete to develop technology. Rich textual form indicates a spherical mapping similar to TNIC.
*Highly volatile data structure, but crucial to our economic future.
Breakthrough innovations Declining?

Technological Disruptiveness (rapidly growing vocab) vs Time

- Many temporary spikes
- They don’t persist!
- Pervasive low frequency decline since WW II
- BFH (2018): this explains part of decline in IPOs
*HPP (2014JF) examines rate of change of vocabulary in firm 10-Ks (product market fluidity).
*Increased fluidity indicates more agile rivals and direct threats.  
*Firms respond by saving liquidity (more cash, fewer dividends).
Textual Network Industries (TNIC) was born on a PC in 2006.
Stage 1 model of U.S. public firms completed in late 2008.
Much completed research done following stage 1.
Stage 2 expansion to U.S. private firms in 3-6 months.
Work under way using same technologies on innovation data.
Implications for optimizing corporate strategy, investors, researchers, regulators and more.