Text-Based Connected Stocks for IPO Firms

Mingjun Sun

Tepper School of Business, Carnegie Mellon University • mingjuns@andrew.cmu.edu

Abstract

How much of an IPO firm's capital structure and returns is explained by textually similar public comparables? I identify connected stocks from the semantic similarity between an issuer's S-1/F-1 business description and seasoned firms' 10-K disclosures. Connected firm leverage predicts IPO leverage beyond industry averages; the returns of connected firms before the IPO predict higher first day returns but lower abnormal returns at one and two years, with offer prices only partially impounding the signal during bookbuilding; and the idiosyncratic volatility of connected firms predicts issuer volatility after listing and pricing uncertainty. Economically similar public firms carry information for the financing and pricing of firms entering public markets.

Motivation

- How should investors value a newly listed firm when little market history is available? In practice, valuation leans on public comparables: underwriters market the offering relative to listed firms, institutional investors anchor to comparables, and issuers frame their business by reference to established firms.
- Standard classifications such as SIC codes are coarse, and manually collected comparables are ad hoc. The economically relevant comparable set is hard to identify.
- I build a granular comparable set for each issuer from public disclosures and ask whether it carries information for IPO financing and pricing, organized around three channels:
  - **Capital structure benchmarking** — comparables reveal financing appropriate for similar assets and risk.
  - **Valuation signal** — returns of comparables before the IPO are an observable signal only partially impounded at the offer.
  - **Benchmark uncertainty** — volatility of comparables proxies for uncertainty about the valuation benchmark.

Data and Sample

- U.S. IPOs, 1996–2025; financial and utility issuers excluded.
- Issuer S-1/F-1 business descriptions and prior year 10-K Item 1 descriptions of public firms, from EDGAR.
- CRSP returns, Compustat fundamentals, and offer price ranges from the registration filings.
- Continuous variables winsorized at the 1st and 99th percentiles; leverage measured at listing.

A Network of Textually Connected Stocks

- Embed each issuer's S-1/F-1 business description and each public firm's prior year 10-K Item 1 description with OpenAI's `text-embedding-3-small` model.
- Measure proximity by cosine similarity; the connected set is the issuer's K closest public firms (baseline $K \in \{10, 30\}$), fixed at the offering.
- The network resolves business similarity at a finer level than SIC codes and captures the comparables observable to market participants at the time of the IPO.

Finding 1: Capital Structure Benchmarking

Connected firm leverage predicts IPO leverage at listing, beyond same SIC2 leverage, issuer characteristics, connected firm fundamentals, and industry and year fixed effects.

Connected firm leverage → IPO leverage	**Book**	**Market**
OLS coefficient (fully controlled)	0.19***	0.22***
IV (2SLS) coefficient	**0.56***	-0.08
First stage F	13.6	22.0

- A 10 pp rise in connected firm book (market) leverage maps to a **1.9 (2.2) pp** rise in issuer book (market) leverage; same SIC2 leverage is small and insignificant.
- Instrumenting connected firm leverage with idiosyncratic equity shocks to large connected firms yields a positive, **marginally significant** book leverage effect ($p = 0.05$ under CR2 standard errors clustered by SIC2 and offer year); market leverage is insignificant.

Finding 2: Valuation Signal and Partial Adjustment

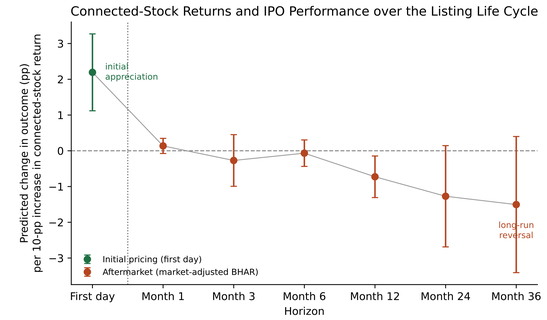
- A 10 pp higher cumulative connected return before the IPO predicts a **2.2 pp** higher first day return; offer prices move only part of the way to the signal.
- Decomposing the move from the initial filing range midpoint to the first day close: the operative information arrives during bookbuilding.

Where the connected return signal is impounded

28.8% offer price revision + **71.2%** first day trading

Finding 3: Reversal at Long Horizons

- The same signal reverses at longer horizons: 10 pp of connected appreciation before the IPO predicts -0.7 **pp** and -1.3 **pp** market adjusted performance at one and two years.
- The reversal loads on the connected stocks themselves, concentrated where strong appreciation coincides with large upward offer price revisions.



Finding 4: Volatility and Price Discovery

- A one standard deviation rise in the idiosyncratic volatility of connected firms forecasts the issuer's own volatility after listing, with coefficients of **0.085**, **0.077**, and **0.059** over days 21–63, 64–126, and 127–252 after listing (each significant at 1%): about **6–9%** higher issuer volatility.
- The same volatility widens the initial filing range (**0.018**, SE 0.008) and enlarges the absolute offer price revision during bookbuilding (**0.011**, SE 0.002), each per one standard deviation.
- It is informative about **risk and price discovery**, not the direction of returns: connected volatility does not robustly predict abnormal returns.

Conclusion

⋆ A network of economically similar public firms, built from public disclosures, carries information for the financing and pricing of IPOs: it predicts capital structure at listing, the partial adjustment of offer prices to a valuation signal, the reversal of that signal at long horizons, and risk after listing. The comparables network is a natural input for settings well beyond the IPO, including seasoned offerings, mergers, and the propagation of firm level shocks across related firms.

Contact

Mingjun Sun • Tepper School of Business, Carnegie Mellon University
Email: mingjuns@andrew.cmu.edu