Insert title about here

Insert authors about here

Abstract

JEL Codes:
Keywords:

Contents

1. Methodology

My initial idea was to take Hu and Härdle (2021), and do it right by doing the NER in a rigorous way using LLMs:

- instead of considering that news articles embed a leader-follower relationship, we don't impose any structure in the news articles
- we perform NER in a more realistic way, by having an LLM parse the news articles and extracting the firms that it considers as "*directly affected by the news articles*". The problen with Hu and Härdle (2021)'s NER is that they need to assume that every firm mentioned in a news article is relevant. This is actually not the case in most news articles, where many firms are mentioned contextually, or even more extreme, sometimes there is no relationship going on between the firms mentioned in the article. For example, we could have a news article like this: "*Moodys lowers the credit rating of Banco Santander*". It's clear that this article is not talking about the existence of a relationship between Moodys and Banco Santander, however, in Hu and Härdle (2021)'s logic, these article defines a connection between those two firms.

However, I am not invested in this idea and would be more than happy to do something different. Below, I deploy some methodology. Again, I am not invested in it, if you want to propose a different methodology, feel free.

1.1 Introduction

The increasing availability of textual data from business news articles provides a rich source of information for studying the relationships between firms. Traditionally, firm networks inferred from such data rely on simple co-occurrence models, where firms are assumed to be connected if they are mentioned together in an article.

In particular,

- Let $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ represent the set of n firms we are analyzing.
- Let $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ represent the dataset of m news articles that mention these firms.

We assume that the news articles can be mapped to specific dates, allowing for a time dimension if needed, i.e., $A_i(t)$, where t represents the publication date of article A_i.

Firm-Article Matrix (Incidence Matrix)

- Define an incidence matrix $M \in \{0,1\}^{n \times m}$, where the entry $M_{ij} = 1$ if firm F_i is mentioned in article A_j, and $M_{ij} = 0$ otherwise.
- This matrix allows us to encode which firms are co-mentioned in the same articles.

Co-occurrence Matrix

From the incidence matrix M, we can construct a co-occurrence matrix $C \in \mathbb{R}^{n \times n}$, where each entry C^{ij} captures the number of articles in which firms F_i and F_j are co-mentioned. Mathematically, this can be expressed as:

$$C = MM^T$$

Here, C^{ij} counts the number of articles that mention both firm F_i and firm F_j.

Weighted Network Representation

The co-occurrence matrix C can be used to define a weighted undirected graph $G = (\mathcal{F}, \mathcal{E}, w)$, where:

- $\mathcal{F}$ is the set of firms (nodes).
- $\mathcal{E} \subseteq \mathcal{F} \times \mathcal{F}$ is the set of edges between firms, where an edge exists between firms Fi and F^j if $C^{ij} > 0$ (i.e., they have been co-mentioned in at least one article).
- $w : \mathcal{E} \to \mathbb{R}+$ is the weight function, where the weight of the edge between firms F_i and Fj is given by $w\left(F_i, F^j\right) = C^{ij}$. This weight represents the strength of the connection between the two firms, based on the number of co-occurrences in news articles.

However, this approach does not account for the nature or type of relationships between firms, nor does it consider the directionality or complexity of those relationships.

In this paper, I propose a novel methodology for constructing firm networks using Large Language Models (LLMs) to analyze the textual content of news articles. The LLM is tasked with two goals: (1) determining whether a substantive relationship exists between a pair of firms based on the context provided in the article, and (2) classifying the type of relationship (e.g., supplier-customer, competitor, partnership). Additionally, I incorporate the **directionality** of certain types

of relationships, such as supplier-customer or mergers and acquisitions (M&A), where relationships are inherently asymmetric. In particular, directionality refers to relationships between different firms where $F_i \to F_j$ but $F_j \not\to F_i$. I also consider **reflexivity**, where firms can have self-relations, such as internal restructuring or stock buybacks. In this case $F_i \leftrightarrow F_i$ The methodology yields a nuanced and comprehensive firm network that captures both the strength and type of relationships between firms, providing a powerful tool for analyzing firm interactions, market dynamics, and the effects of external shocks.

1.2 Mathematical Framework for Firm Networks Using LLMs

1.3 Firm Set and News Articles

Let $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ represent the set of n firms under consideration. Each firm is potentially mentioned in a set of news articles $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$, where m denotes the number of articles. Each article $A_i \in \mathcal{A}$ contains textual content $T(A_i)$ and is published on a specific date $t(A_i)$.

1.4 LLM-based Relationship Detection and Classification

For each article A_i, the LLM processes the textual content $T(A_i)$ and performs two key tasks:

1. **Relationship Detection**: The LLM determines whether there is a substantive relationship between a pair of firms $(F_i, F_j) \in \mathcal{F} \times \mathcal{F}$ based on the context provided in article A_k.

2. **Relationship Classification**: If a relationship exists, the LLM classifies the relationship between (F_i, F_j) described in article A_k into a relationship type $r_{ij}(A_k) \in \mathcal{T}$, where $\mathcal{T} =$ {Supplier, Competitor, Partnership, M&A, Legal, Other} is the set of relationship types. Note that some relationships are "*directional*", while others are not. In particular:

   - *Supplier-Customer*: $F_i \to F_j$, where F_i is the supplier and F_j is the customer.
   - *Competitor*: $F_i \leftrightarrow F_j$, where both firms compete for market share.
   - *Partnership*: $F_i \leftrightarrow F_j$, where the firms collaborate on a project or initiative.
   - *Mergers & Acquisitions (M&A)*: $F_i \to F_j$, where firm F_i absorbs or acquires firm F_j.
   - *Legal Dispute*: $F_i \to F_j$, where firm F_i sues or takes legal action against firm F_j.
   - *Other*: contains the rest of relationships that the LLM was unable to clasify in the previous categories. For simplicity, we make this an undirected relationship, so $F_i \leftrightarrow F_j$.

Note that in these defitions, order matters, as we are always considering that $F_i \to F_j$ in any directional relationship between any $(F_i, F_j) \in \mathcal{F} \times \mathcal{F}$.

The LLM also assigns a relationship score LLM_score(A_k, F_i, F_j, r), reflecting the confidence in the existence and strength of the relationship $r_{ij}(A_k)$ between firms F_i and F_j based on article A_k.

1.5 General Relationship Matrix

Let $\mathcal{R}(A_i) \subseteq \mathcal{F} \times \mathcal{F}$ represent the set of firm pairs (F_i, F_j) that the LLM determines to be related based on the content of article A_i. The task of the LLM is to analyze the article $T(A_i)$ and determine when a meaningful relationship or event connects the firms.

For each article A_i, the LLM processes the text $T(A_i)$ and returns a set of firm relationships:

$$\mathcal{R}(A_i) = \left\{ (F_i, F_j) \in \mathcal{F} \times \mathcal{F} \mid \text{LLM concludes } F_i \text{ and } F_j \text{ are economically tied in } A_i \right\}$$

The key here is that $\mathcal{R}(A_i)$ is determined by the LLM's understanding of the text, identifying cases where firms are tied by contracts, joint ventures, lawsuits, partnerships, or other significant business events, rather than simple co-mentioning.

From the set of firm relationships across all articles, we can construct a relationship matrix $R \in \mathbb{R}^{n \times n}$, where each entry R_{ij} quantifies the strength of the relationship between firm F_i and firm F_j. The entry R_{ij} is computed as:

$$R_{ij} = \sum_{k=1}^{m} \mathbb{1}_{\{(F_i, F_j) \in \mathcal{R}(A_k)\}} \cdot \text{LLM\_score}\left(A_k, F_i, F_j\right)$$

1.6 Type-Specific Relationship Matrix

Since we have richer information about the relationship type, we can define a relationship matrix that is specific to each type of relationship $r \in \mathcal{T}$. For each article A_i, we now have:

$$\mathcal{R}(A_i) = \left\{ \left(F_i, F_j, r_{ij}(A_k)\right) \in \mathcal{F} \times \mathcal{F} \times \mathcal{T} \;\middle|\; \text{LLM detects and classifies a relationship} \right\}$$

For each relationship type $r \in \mathcal{T}$, we define a relationship matrix $R^r \in \mathbb{R}^{n \times n}$, where the entry R^r_{ij} quantifies the strength of the relationship r between firms F_i and F_j.

$$R^r_{ij} = \sum_{k=1}^{m} \mathbb{1}_{\{(F_i, F_j, r_{ij}(A_k)) \in \mathcal{R}(A_k)\}} \cdot \text{LLM\_score}(A_k, F_i, F_j, r_{ij}(A_k)),$$

where:

- $\mathbb{1}_{\{(F_i,F_j,r_{ij}(A_k))\in\mathcal{R}(A_k)\}}$ is an indicator function that equals 1 if article A_k identifies relationship r between firms F_i and F_j, and 0 otherwise.

- $\text{LLM\_score}(A_k, F_i, F_j, r_{ij}(A_k))$ is the score provided by the LLM that quantifies the strength of the relationship.

For some relationship types, such as *supplier-customer*, the matrix R^r is **asymmetric**, meaning $R^r_{ij} \neq R^r_{ji}$. In contrast, for *competitor* or *partnership* relationships, the matrix R^r is **symmetric**, meaning $R^r_{ij} = R^r_{ji}$.

1.7 Reflexivity in Firm Relationships

In addition to relationships between firms, we also consider **reflexive relationships**, where a firm F_i has a relationship with itself, denoted $F_i \leftrightarrow F_i$. Reflexivity can capture internal actions such as:

- *Firm Restructuring*: Internal reorganization or governance changes.

- *Stock Buybacks*: Financial actions where a firm repurchases its own shares.

- *Internal Legal Actions*: Actions that affect a firm's own internal compliance or governance.

The reflexive relationships are represented in the diagonal elements R^r_{ii} of the relationship matrix for each type r:

$$R^r_{ii} = \sum_{k=1}^{m} \mathbb{1}_{\{(F_i,F_j,r_{ij}(A_k))\in\mathcal{R}(A_k)\}} \cdot \text{LLM\_score}(A_k, F_i, F_i, r_{ij}(A_k)),$$

where the diagonal element R^r_{ii} quantifies the strength of firm F_i's reflexive relationship under relationship type r.

Specifically:

- If $R^r_{ii} > 0$, firm F_i has a reflexive relationship in the context of relationship type r (e.g., self-influence or self-reference).

- If $R^r_{ii} = 0$, firm F_i does not have a reflexive relationship.

1.8 Network Representation with Directionality and Reflexivity

The firm network is constructed as a **multi-layered graph** $G = \{G^r\}_{r \in \mathcal{T}}$ composed of relationship-specific layers $G^r = (\mathcal{F}, \mathcal{E}^r, w^r)$, where:

- $\mathcal{F}$ is the set of firms (nodes).
- $\mathcal{E}^r \subseteq \mathcal{F} \times \mathcal{F}$ is the set of edges representing relationships of type r, where an edge exists between F_i and F_j if $R^r_{ij} > 0$. Depending on the type of relationship $r \in \mathcal{T}$, the edges may be **directed**, **undirected** or **looping**:
  - *Directed edges*: For relationships such as supplier-customer, M&A, and legal disputes, the edges are directed, representing asymmetric relationships where $F_i \to F_j$ but not necessarily $F_j \to F_i$.
  - *Undirected edges*: For symmetric relationships like competition and partnerships, the edges are undirected, meaning $F_i \leftrightarrow F_j$.
  - *Looping edges*: For reflexive relationships where $F_i \leftrightarrow F_i$. The weight of the self-loop reflects the strength of the firm's internal actions.
- $w^r : \mathcal{E}^r \to \mathbb{R}_+$ is the weight function, where the weight of the edge between F_i and F_j is given by $w^r(F_i, F_j) = R^r_{ij}$, representing the strength of relationship type r between the two firms.

This creates multiple layers of networks, each representing a different type of firm interaction.

1.9 Dynamic Networks and Temporal Analysis

To capture how relationships between firms evolve over time, we introduce a **time-varying relationship matrix** for each relationship type r:

$$R^r_{ij}(t) = \sum_{k=1}^{m} \mathbb{1}_{\{(F_i, F_j, r_{ij}(A_k)) \in \mathcal{R}(A_k)\}} \cdot \text{LLM\_score}(A_k, F_i, F_j, r_{ij}(A_k)) \cdot \mathbb{1}_{\{t(A_k) = t\}}.$$

This matrix captures the relationships at a specific time t based on the articles published during that time period. The resulting **dynamic network** $G(t)$ allows for the study of how firm interactions and network structures evolve in response to economic events, mergers, or external shocks.

1.10 Analytical Methods and Potential Insights

With the constructed network, several analyses can be performed to extract valuable insights:

- **Centrality Measures**: Compute various centrality measures (degree, eigenvector, betweenness) to identify key firms within specific relationship networks. For example, firms central in the "supplier" network might play crucial roles in supply chains, while those central in the "competitor" network might dominate their industries.
- **Community Detection**: Apply community detection algorithms to identify clusters of firms that are closely related. Different clusters may emerge in different layers, such as a supply chain ecosystem or a competitive industry cluster.
- **Temporal Analysis**: Track the evolution of firm relationships over time, analyzing how major economic events (e.g., financial crises, regulatory changes) impact the structure and strength of firm networks.
- **Impact of Reflexivity**: Study firms with significant self-relations (high diagonal values in the relationship matrix) to understand how internal actions, such as restructuring or stock buybacks, affect firm performance and market position.

References

J. Hu and W. K. Härdle. Networks of news and cross-sectional returns. *arXiv preprint arXiv:2108.05721*, 2021.