



Final Session

Session 6

Thomas Rigou

Group Project

- **Goal** : Replicate and extend the paper by Amaya and Filbien (2015) on ECB's Communication.
 - Extensions can target the textual analysis methodology (similarity, sentiment, preprocessing...) as well as the exogenous variables used, regressions performed...
 - Extensions should be well argumentated.
- **Report** : 5 pages max (Not including Appendix or References)
- **Code** : Folder with all code and data (Github, Google Drive...)..
I should be able to retrieve your main figures and tables from the code, and run it with minimal edit.
- **Oral presentation** : 10/15 min + Individual Q&A

Text as Data : Key Decisions

Unit of Analysis

- Whole document
- Paragraph
- Sentence
- Word / token

Informative Text Parts

- All text
- Specific sections
- Specific Speakers
- Keyword-centered text

Information Types

- All words
- Content words (Noun, Verb, Adjective)
- Named entities
- Numbers & symbols

Preprocessing Choices

- Lowercasing
- Stopwords
- Stemming / Lemmatization
- Punctuation & boundaries

Text Representation

- BoW / TF-IDF
- Topic models
- Lexicons
- Embeddings / LLMs

Evaluation Constraints

- Interpretability
- Replicability
- Computation
- Theory alignment

Every preprocessing decision is a modeling assumption.

If you cannot justify it, it probably should not be there.

Risk and Uncertainty Indexes

- **Measurement Challenge:**
Because uncertainty cannot be directly observed, empirical research must rely on proxies that capture agents' perceptions rather than realized outcomes.
- **From Volatility to Perceptions:**
Early measures focus on market or macro volatility, which reflect realized shocks but do not distinguish between expected and unexpected risk or between different sources of uncertainty.
- **Text-Based Uncertainty/Risk Indexes:**
A prominent literature uses language as a proxy for perceived uncertainty, based on the idea that economic agents reveal concerns about risk and uncertainty in written and oral communication.

Hoberg and Phillips : Text-based Network Industry Classifications

- **Limitations of Traditional Industry Classifications:**
Standard classifications such as SIC or NAICS are static, infrequently updated, and often group together firms with very different products, while failing to capture new markets and innovation-driven changes.
- **Transitivity and Homogeneity Assumptions:**
Traditional industry codes impose the assumption that competition is transitive and homogeneous within industries, even though firms may compete with some rivals but not others within the same category.
- **Lack of Competitive Intensity and Distance:**
Existing classifications provide no notion of how close or distant firms are, treating all competitors within an industry as equally similar.

Hoberg and Phillips : Text-based Network Industry Classifications

- What Unit of Analysis ?
 - Firm-Level Analysis -> Full Document
- Which Text section to use ?
 - Item 1 Business Description; more specifically Product Description section
- Which Information Types ?
 - Only Noun and Proper Nouns (words that are capitalized 90%+ of the time)
- Preprocessing choices ?
 - Drop words appearing in 25%+ of firm's descriptions (firm, company, business...)
 - Drop country, state, and city names