The Knowledge Graph for Macroeconomic Analysis with Alternative Big Data

Yucheng Yang

Joint with Yue Pang (PKU), Guanhua Huang (USTC) and Weinan E (Princeton)

November 11, 2020

2020 Banca d'Italia and Federal Reserve Board Joint Conference

Motivation

- Traditional macroeconomic models only have a handful of variables.
- Big data and machine learning allows us to develop models with much more variables.
- Most papers put large number of variables into statistical models (nowcasting, factor model, etc.) directly, without understanding their relationships.
- We need a new knowledge system on relations among traditional and many new economic variables to design model inputs.
- This paper: we build a knowledge graph (KG) of the linkages between traditional and alternative data variables.

Introduction: Knowledge Graph

- Knowledge graph: knowledge base that uses graph topology to represent interlinked descriptions of entities.
- Basic elements: “RDF triple” with form {subject, predicate, object}.
- Prominent application: Google Search.

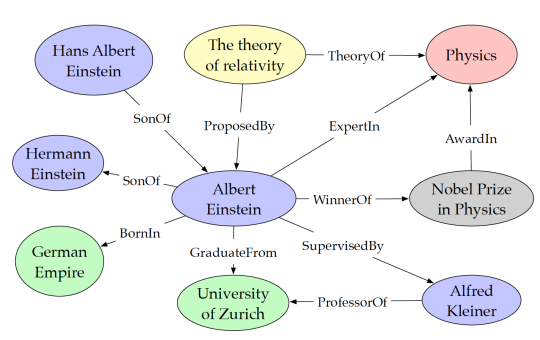
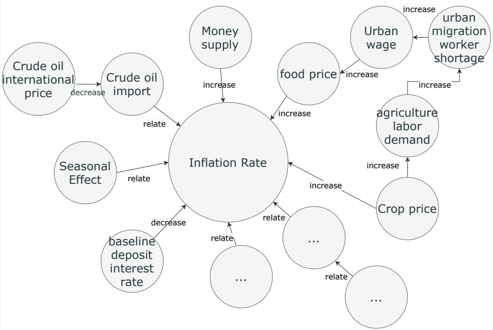


Figure 1: Example of Knowledge Graph on Einstein (Ji et al., 2020)

This Paper



- We build a knowledge graph (KG) of the linkages between traditional economic variables and alternative data variables.
- The "RDF triples" are extracted from academic literature and industry research reports.
- We apply the knowledge graph of economic variables to do variable selection in economic forecasting.

Textual Data for Knowledge Graph Construction

- Data: industry macro research reports from China.
  1. Focus on analyzing or forecasting the dynamics of aggregate variables, and it is always clearly stated what variables are studied in each report.
  2. They mostly adopt the narrative approach (Shiller, 2017), which clearly state the logic chains of their analysis in narrative language, rather than in theoretical or quantitative models.
  3. Freely available and can be downloaded massively from the WIND database

Construction of Knowledge Graph: An Example

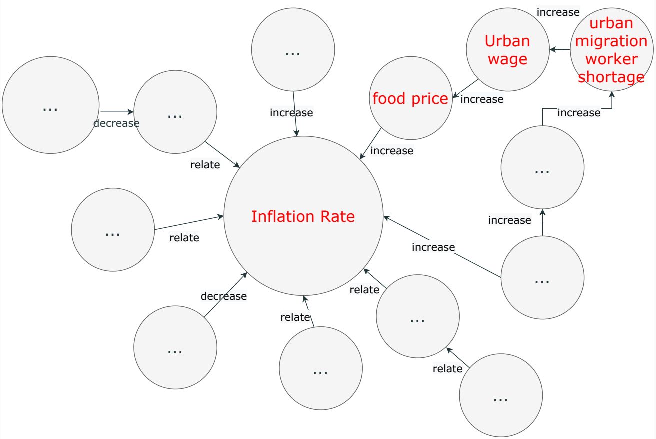
> "A long-term systematic **migrant worker shortage** began to appear in the Chinese migrant labor market around 2005, which greatly **increased** the **growth rate of migrant workers' wages**, **resulted in the increase** of **food prices**, and **pushed up** the increase in **consumer price index**, **making** the average level of **inflation** probably 100 to 200 basis points **higher**."

RDF triples of {variable 1, relation, variable 2} format:

- {migrant worker shortage, increase, growth rate of migrant workers' wages}
- {growth rate of migrant workers' wages, resulted in the increase, food prices}
- {food prices, push up, consumer price index}
- {food prices, make higher, inflation}

Construction of Knowledge Graph: An Example

After removing duplicates, we get:



Main Challenges: Entity Recognition

Entity Recognition is very hard, since economic variables are mostly multi-token entities with complicated semantic patterns.

- Examples: “migration worker shortage”, “growth rate of migration workers’ wages”, “processing firm registrations in China”, “leverage rate of local government financing vehicles”.

We develop a weakly supervised learning algorithm with human involvement to extract variable entities and relation keywords from the textual data.

Construction of Knowledge Graph: Results

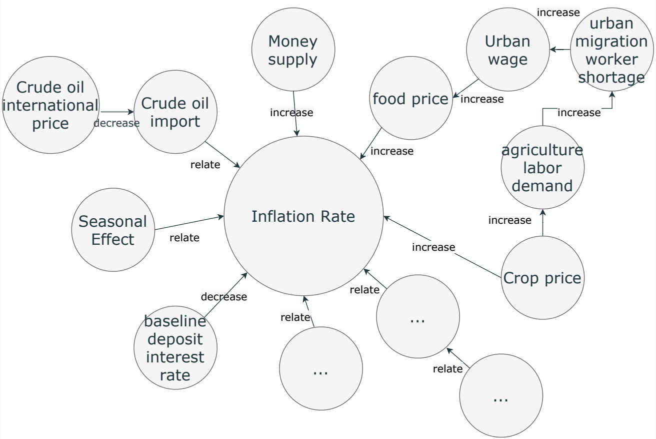


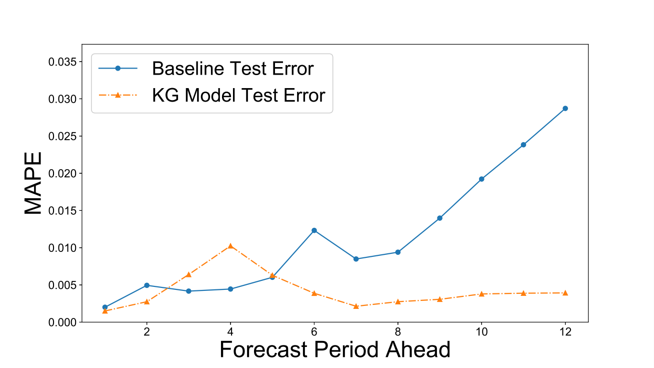
Figure 2: Example of Knowledge Graph on Inflation

Application of Knowledge Graph: Economic Forecasting with Many Inputs

We forecast China's monthly *inflation rate* and *nominal investment* time series from April 1996 to June 2019.

- Baseline model: standard time series constructed by Higgins and Zha (2015) as model inputs + statistical method (Lasso).
- KG-based model: model inputs guided by the knowledge graph + Lasso.

Application of Knowledge Graph: Inflation Forecasting



- **Short run**: forecast errors for both models are comparable. Baseline model even outperforms KG-based model in some horizons.
- **Long run**: baseline model gets worse, while the KG-based model achieves a stable and much higher accuracy.
- **Test of comparison**: comparisons are significant under Diebold-Mariano test.
- "Short term forecasts rely on statistics, long term on logic." KG could better capture underlined logic of the economy than statistical methods on big data.

Conclusion

- In age of big data, macroeconomics need a new knowledge system with more economic variables.
- We develop an approach to build a knowledge graph (KG) of the linkages between traditional and alternative data variables from textual data.
- We apply the KG to variable selection in economic forecasting.
- Compared to statistical methods, KG-based method achieves higher forecasting accuracy, especially for long term forecasts.
- Many other exciting applications on the way!

Link to the Paper

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3707964

Thanks for your attention!