Enterprise OKN: A Federated Knowledge Graph for Financial Data

Jay Pujara
jpujara@usc.edu

Louiqa Raschid
louiqa@umiacs.umd.edu

Gerard Hoberg
hoberg@marshall.usc.edu

Gordon Phillips
Gordon.M.Phillips@tuck.dartmouth.edu

Craig Knoblock
knoblock@isi.edu

The engines of commerce and industry continuously generate streams of data that reflect financial and economic activity. Unfortunately, this rich data is often not captured or curated in machine understandable form, or readily integrated across resources and data streams, presenting an obstacle for research, policy and industry use. We proposed to develop the Enterprise Open Knowledge Network (Enterprise OKN), envisioned to be a shared resource of curated knowledge from heterogeneous sources, together with tools to support large-scale data analysis, and interfaces to allow access to additional repositories. Enterprise OKN, with appropriate ontological knowledge, will create unprecedented opportunities for financial and socio-economic research, will inform data-driven fiscal and economic policy, and will empower innovators and entrepreneurs.

Enterprise OKN will leverage a wealth of unstructured data on the Web about companies and their products and services. It will also exploit semi-structured data and time series datasets, in repositories of the deep Web including data.gov, that have been provided for regulatory or legal purposes. Finally, there are reference datasets, e.g., catalogs and indices, that provide standard identifiers and metadata, that enable cross-resource federation. These data resources inspire several of the technical capabilities of Enterprise OKN. The intellectual challenges of Enterprise OKN include the following:

- Ontologies, entity resolution and linking: Creating the standards for representing financial and economic knowledge. Defining and extending unique identifiers for financial entities and semi-automatically mapping entity and relationship mentions, in textual descriptions, to identifiers.

- Semantic and Temporal Embeddings: Reusable, high-dimensional vector space embeddings from unstructured collections, e.g., company Web pages, patents, and product information for financial entities, their relationships (competitors, joint ventures, supply chain, etc.). Temporal embeddings will reflect temporal evolution, as well as features extracted from time series datasets.

- A suite of tools for community detection, understanding behavior in markets, visual analytics, prediction and outlier detection, etc. customized and tuned to Enterprise OKN.

We motivate Enterprise OKN with use cases around innovation, entrepreneurship, and fiscal policy. A first use case explores the impact of research, investment, and discovery on the process of innovation.We consider the efficacy and speed at which patented technologies and technology transfer impact the product life cycle. Some technologies can significantly and quickly change markets and business processes. Related outcomes include job creation as well as economic booms and busts, such as the technology boom/bust of the 1990s. This stream of research will require reduced-dimension spatial models of discoveries in the technology space, and new product evolution models in the space of product offerings. Text from patent descriptions and claims, mapped to a reduced dimension embedding (doc2vec), can facilitate the former. Text from company websites obtained from the Wayback Machine is one option to monitor the emergence of product offerings, for the latter. A second use case focuses on the opportunities for entrepreneurs to exploit Enterprise OKN. We envision a portal where a prospective small business owner can copy-paste a business plan. She would receive an analysis about the competitive landscape of companies and products. Further, she could be given recommendations of complementary businesses, relevant joint venture opportunities, etc. Such a portal could also benefit the management of licenses and permits, as well as the regulation of competition (antitrust), technology property rights (trademarks and patents), etc. A final use case addresses better monitoring and improved fiscal policy. The 2008 US financial crisis revealed the lack of open data and tools to better understand risks to interconnected financial markets. Soaring values of corporate debt, combined with high levels of stock buyback, is feared to lead to a financial bubble that could trigger the next US recession. Enterprise OKN could provide data and tools for improved monitoring of financial markets. The broader impact of Enterprise OKN: Researchers can re-purpose datasets and use machine learning approaches to address financial and socio-economic research questions at scale. Analysts and regulators can use Enterprise OKN to make more informed policy recommendations. Small businesses will be able to identify partners and competitors. On the educational frontier, a new generation of scholars will blend computational solutions with theories, models and methodologies from finance, economics, mathematics and statistics.

We briefly highlight several recent results that contribute to the development of the Enterprise OKN. These include the Record Linkage Toolkit, which is a package for large-scale entity resolution, WTNIC which is a network of financial competitor relationships for public firms and startups constructed using 20 years of Internet Archive webpages for over 600K companies, and a topic model for community-detection in toxic mortgage-backed securities in the 2009 housing crisis.