Text-Based Nonobvious Peer Group Identification Project Proposal

**Dong Sun**

1. **Overview**

As a baseline, I plan to implement the publication (Hoberg and Pillips, 2017) about testing text-based industry momentum. In the paper, the authors applied cosine similarity (Hoberg and Pillips, 2016) over text vectors to model similarity between firms. I also plan to try out some new models like [Latent](https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation) Dirichlet Allocation (LDA) to capture the latent representation of firms’ business description and build a different similarity measure to discover nonobvious economic links.

1. **Literature review**

The authors used 10-K business descriptions to build pairwise word similarity matrix for pairs of firms (Hoberg and Pillips, 2016). They further claimed that an industry classification model (TNIC) based on the word similarity matrix gives different and informative results compared to traditional SIC-3 and NAICS4 classification rules. In a more recent study (Hoberg and Pillips, 2017), the authors showed that nonobvious economic links discovered by TNIC provide a stronger industry momentum variable, compared to own-firm variables and traditional SIC-based peer variables.

1. **Methodology**

There are three main steps included in this proposal: First, we will parse public traded firms’ business

description from the SEC website and store them into a database. Second, we will perform text preprocessing

and use language models like LDA to get a similarity matrix of the all the firms. In addition, we need to build an

classification rule based on the similarity matrix. Finally, we will use the classification result to build and test an industry

momentum factor, using Fama-MacBeth regression.

* **Data Collecting**:

***Step1***: First we can collect publicly traded firm names in the CRSP database. After that, we gather business descriptions from these firms annual 10-Ks filed (From Item 1 or Item 1A sections) with the SEC using web crawling algorithms in Python. Python is useful for creating web crawlers. Some libraries including *requests*, *selenium* can be used to make crawling request; libraries like *beautifulsoup* are useful for analyzing crawled webpages.

***Step2***: Create a database to save information like SIC code, company name and company description for further analysis.

* **Text Modelling:**

***Step3***: For text preprocessing, the authors only picked nouns and uncommon pronouns from the business

description. We can use this method as a starting point.

***Step4***: First we will convert the preprocessed company descriptions into bag of words/tf-idf representation. After that, I will run LDA using those representations and get a vector representation for every firm. Finally, we will calculate the similarity matrix between vector representations using some distribution similarity measures like Jensen-Shannon Distance. Besides LDA, some other advanced topic models like lda2vec is also worth trying. Python library likes *NLTK* and *genism* are useful for text analysis.

* **Hypothesis Testing:**

***Step5***: After calculating the similarity matrix, we can use thresholds to identify industry groups. In the paper, the authors consider firm pairs with similarity score larger than 0.2132 to be in one industry group. In order to test the first hypothesis, we can build the industry momentum factor by calculating equal-weighted lagged returns from a firm’s industry peers. By regressing the own firm’s return over industry momentum factor and some other control variables, we are able to test the statistical strength of the industry momentum factor. Python library *statsmodels* are useful for hypothesis testing.

1. **References**

* Hoberg, G. & Phillips, G. M. Text-based industry momentum (2017)
* Hoberg, G. & Phillips, G. M. Text-based network industries and endogenous product differentiation (2016)