**Project Report**

Group name: CLZZ

Students: Anhua Chen (anhua@uchicago.edu)

Xiang Zhang (snzhang@uchicago.edu)

Xiuyuan Zhang (xiuyuanzhang@uchicago.edu)

David Liu (dliu5@uchicago.edu)

Github repository: <https://github.com/zhangxiang0822/CS123_Project>

**Project Overview**

This project explored the topic of innovation diffusion across industries using patent citation data. Through analyzing the dynamics of the patent citation behaviors across industries, we aimed to find the diffusion patterns of patents within and across industries. To achieve this goal, we performed several large data merges leveraging the Google Cloud Platform for parallel computing; further, we used the Dijkstra algorithm to find the shortest paths between each patent citation pair and aggregated these pairwise computations to the industry level. For this specific project, we focused on patents granted in the U.S due to the magnitude of the citation data. This means that, for a citation pair where A cited B, both A and B were patents granted in the U.S.

**Data Description**

We used a total of three datasets: (1) NBER patent classification data, (2) U.S. patent citation data, and (3) NBER industry classification data. All three datasets were downloaded through the United States Patent and Trademark office website: <http://www.patentsview.org/download/>

**(1) U.S. patent citation data** (98,207,057 rows, 3.505 GB). The U.S. patent citation data provides information on citations of U.S. patents made by U.S. patents. Each row of the dataset contains nine columns: universally unique id, patent number, patent number for which the current patent cites, date of when the cited patent was granted, name of the cited record, kind code from WIPO, country where the cited patent was granted (always U.S.), category (who cited the patent), sequence (order in which this reference is cited by select patent).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| uuid | patent\_id | citation\_id | date | name | kind | country | category | sequence |
| 00000jd7thmiucpaol1hm1835 | 5354551 | 4875247 | 1989-10-01 | Berg | NULL | US | NULL | 11 |
| 00000l0ooxrvfv6jkenobhwis | D674253 | D519335 | 2006-04-01 | Ishii | S | US | cited by examiner | 13 |
| 00001nlwuimui60vu3k1yzjqd | 8683318 | 6642945 | 2003-11-01 | Sharpe | B1 | US | cited by examiner | 6 |
| … | … | … | … | … | … | … | … | … |

Table. 1 U.S. patent citation sample data

**(2) NBER patent classification data** (5,105,937 rows, 110.807 MB).Each row of the NBER patent classification data contains category information for a single patent, including the universally unique identifier for the patent, the patent number, a NBER category id, and a NBER subcategory id.

|  |  |
| --- | --- |
| id | name |
| 11 | Agriculture, Food, and Textiles |
| 12 | Coating |
| 13 | Gas |
| … | … |

|  |  |  |  |
| --- | --- | --- | --- |
| uuid | patent\_id | category\_id | subcategory\_id |
| 000114qfli99qqd9fsbxichy1 | 6243839 | 2 | 22 |
| 0001jsdl1xi7z84rzx9iwvdlh | 4646100 | 2 | 21 |
| 0001qpsb0yts8daudtuf3mbm8 | 7712627 | 6 | 68 |
| … | … | … | … |

**(3) NBER industry subcategory data** (37 rows, 871 bytes). This data contains information of 37 pairs of subcategory IDs and subcategory names. The set of subcategories is developed by the National Bureau of Economic Research (NBER) and widely used for economic research.

Table. 2 NBER patent classification

sample data

Table.3 NBER industry subcategory sample data

**Data Merge & Preprocessing**

**Summary Statistics**

**Finding the Shortest Path**

**Discussion**

**Reference**