CS 838 - Project
Stage 1 - Define Problem & Collect Data

Group ID: 1
Group members: Trang Ho; Thomas Ngo; Qinyuan Sun

1 - QUESTION SET

This project aims to examine **how researchers' mobility affects their post-mobility collaboration at their hiring organization**. This contributes to the literature of human capital mobility, which has argued that employee mobility enables organizations to access new knowledge and resources - determinants of organizations' competitive advantage. Such benefit of hiring new employees is conditional on not only human capital and social capital of the new employees but also to what extent their human capital and social capital is disseminated within hiring organizations. For collaboration is an important way for individuals to share their human capital and social capital, we are interested in the outcome factor of post-mobility collaboration of new researchers at the their hiring organization. Drawing on human capital theory, we hypothesize that:

> *Researcher **mobility** is positively associated with the number of **post-mobility collaboration** between new employees and incumbent employees at hiring organizations*

Researcher mobility is measured by a dummy variable indicating whether a researcher moved from one to another affiliation in a particular year. For example, if researcher X was in affiliation A in 2010 and in affiliation B in 2011. We consider that his mobility happened in 2011.

Post-mobility collaboration is measured by the number of papers coauthored by a focal researcher and researcher(s) at the hiring affiliation.

2 - DATASET

In order to test our hypotheses, we use the following empirical data:

- Context of management researchers, whose data comes from the Academy of Management (AOM) conference website (http://aom.org/about/).
- Affiliation information, which data comes from International Association of Universities (IAU) website (http://www.iau-aiu.net/)

2.1 - AOM conference website

Academy of Management is the premier international association for researchers and professionals in the field of management and organization. Founded in 1936, the association has recruited about 20,000 members from about 120 countries and has organized 76 annual conferences. For each annual conference, researchers submit papers for peer reviews and, AOM conference data contains information on those papers accepted as a result of the reviews.

As of now, we have AOM conference data for 3 years, from 2010 to 2012; and each year has approximately 8000 individuals and 4000 papers. In total, we have about 32,000 observations of year-individuals and 16,000 observations of year-papers. The data extracted are put in separate csv files for each year under this folder.

Below are the data structures extracted from AOM:

Table 1 - Researcher Data from AOM Conference Website

#	Attributes	Description
1	Person ID	Unique ID for each individual within a year and (presumably) across years
2	Name	Researcher's name
3	Affiliation	Researcher's affiliation
4	Country	Researcher's country
5	State	Researcher's state
6	City	Researcher's city
7	Phone number	Researcher's phone number
8	Fax number	Researcher's fax number
9	Email address	Researcher's email address
10	Paper title	Papers presented by researchers

Table 2 - Paper Data from AOM Conference Website

#	Attributes	Description
1	Session ID	Unique ID of the session where a paper was included
2	Session search code	Keywords of the session (these keywords were chosen by AOM organizers and they are supposed to indicate the theme of each session)
3	Paper title	Paper's title
4	Paper author's name	Names of authors for each paper
5	Paper author's affiliation	Affiliations of authors for each paper
6	Paper abstract	Abstract of paper
7	Paper search code	Keywords of the paper (these keywords were chosen by authors and they are supposed to indicate the main topic of each paper)

2.2 - International Association of Universities website

The International Association of Universities website provides list of all education affiliations worldwide with detailed information such as postal code (for spatial proximity calculation), country, etc. Because affiliation names may vary across individuals in AOM (e.g: "University of Wisconsin, Madison" can be "*University of Wisconsin-Madison*", "*University of Wisconsin - Madison*", "*University of Wisconsin Madison*", or "*UW-Madison*"), we will use affiliations extracted from IAU as standard entities list to match with affiliation names from AOM. Right now, we have extracted about 17,500 affiliations from IAU. The csv file is stored in this folder.

Below is the data structure of affiliation extracted from IAU

Table 3 - Affiliation Data from IAU Website

#	Attributes	Description
1	Affiliation ID	Unique ID for each affiliation
2	Affiliation name	Name of the affiliation
3	Affiliation country	Country of the affiliation
4	Affiliation street	Street address of the affiliation
5	Affiliation city	City of the affiliation
6	Affiliation province	Province of the affiliation
7	Affiliation postal code	Postal code of the affiliation
8	Affiliation website	Website URL of the affiliation

2.3 - Data Extraction Tools

To extract the structured data, we implemented a Python script using the open-source library *BeautifulSoup*[1] to scrape data from the two website AOM and IAU.

3 - TEXT DOCUMENTS

We want to extract affiliation (university) names from a text documents. Hence, we will search for keyword "*university*" from the *U.S. section* on The New York Times, and use resulted articles as text documents. The text documents are stored in this folder. Below is an example of a text document. The example contains 2 university names - the University of California, and the University of Washington.

Document example:

[1] Beautiful Soup is a Python library which allows to parse HTML page, then navigate, search, or extract needed data.

"The videotaped sucker punch that staggered the white nationalist Richard Spencer on Inauguration Day quickly inspired mockery on social media. But it echoed loudly in an escalating confrontation between extreme ends of the political spectrum. With far-right groups edging into the mainstream with the rise of President Trump, self-described anti-fascists and anarchists are vowing to confront them at every turn, and by any means necessary — including violence. In Berkeley, Calif., on Wednesday night, masked protesters set fires, smashed windows and stormed buildings on the campus of ** the University of California ** to shut down a speech by Milo Yiannopoulos, an inflammatory Breitbart News editor and a right-wing provocateur already barred from Twitter. Five people were injured, administrators canceled the event, and the university police locked down the campus for hours. That followed a bloody melee in Seattle on Inauguration Day, Jan. 20, when black-clad demonstrators — their faces concealed to minimize the risk of arrest — tried to prevent a speech by Mr. Yiannopoulos at ** the University of Washington **, and a 34-year-old anti-fascist was shot and seriously wounded by a supporter of Mr. Yiannopoulos."

4 - LINKS

- Project Website: https://trangho.github.io/cs838-spring2017/
- AOM data: https://github.com/TrangHo/cs838-spring2017/tree/master/dataset/AOM
- IAU data: https://github.com/TrangHo/cs838-spring2017/tree/master/dataset/IAU
- Text documents: https://github.com/TrangHo/cs838-spring2017/tree/master/dataset/text_documents