**Intruction to the Data collection/crawling for Software-as-a-Service (SaaS) Research Project**

This project is to crawl and code data from multiple sources for the research on the performance effect of transitioning to SaaS on legacy on-premise software vendors. There are two parts of the project:

1. collect and code data of “organizational restructuring and reconfiguration” from financial reports, SEC 10-K, and news announcements: and
2. crawl data from Linkedin and code the data to measure the human capital flow among software companies. (note: human capital refers to employees who have SaaS skills)

We might not be able to crawl Linkedin data because Linkedin does not allow the third party to use Bot technology to crawl user profiles from their website automatically. In this case, we may crawl the data from the job announcement database such as Indeed or Monster.

Part (a):

* I will provide a list of the company names (around 1200). Please crawl these firms’ documents from 1998-2018.
* Crawl these firms’ finanical reports and 10-K from EDGAR database. Please refer to Garcia et al. (2012) and Ashraf (2017) for the detail instruction of how to crawl EDGAR. Also there is another paper talking about how to crawl data from EDGAR:

<https://www.hindawi.com/journals/tswj/2014/506740/>

* Crawl these firms’ news announcements from Factiva. Please check the scripts on the Github: <https://github.com/sfu-discourse-lab/SFU_Comment_Extractor>.
* Refer to Girod and Whittington (2017) for the data coding methods to code the downloaded documents (note: I will provide the coding scheme after we finish the document crawling).

Part(b) – after we finish Part(a), we will discuss how to do this part.