# Abstract

For this final project, this paper will explore if Neo4J is a useful and scalable tool for analysis in Institutional Research. Institutional research is broad field that aims to understand how colleges and universities operate (from admissions and financial aid to faculty productivity and research impact), and leverage university data to aid in policy and decision-making.

I work in Harvard’s Office for Institutional Research, and was intrigued to see if Neo4J offered value to some of the analysis work my office does. In class, we only took a superficial look at Neo4J and created a toy database with a couple of nodes. This paper will explore if Neo4J is scalable to larger datasets and if it offers any novel insights.

Due to sensitive nature of the data we work with, I chose to use a publicly available dataset provided by AMiner, the Open Science Platform, which collects Computer Science publications, their citations, and other useful metadata. It is a large dataset with approximate 2.3 million publications and their associated citation data and metadata.

This paper documents my attempts load the AMiner dataset into Neo4J. Next it documents how I used R to analyze the citation-network using RNeo4J, R, RCharts and the Cypher query language. Finally it covers how I developed an RShiny application to showcase my visualizations and analysis results and the deployment of the application on shinyapps.io.

* Working Demo can be found here:
  + <https://certainentropy.shinyapps.io/CS63_project/>
* Source Code is on github:
  + <https://github.com/ddalal2/cs63_final_project>
* Youtbube recordings can be found here:
  + Long: <https://www.youtube.com/watch?v=zc4-wQmv4Pc>
  + Short: <https://www.youtube.com/watch?v=CznumpkD6_M&feature=youtu.be>