

Big Data Systems
Project #3 – Graph Databases/Neo4j
Due Date: June 8th, 2024

In this project you will be given a dataset and you will represent as a graph database in Neo4j, run some queries and measure performance.

1. Go to Stanford Large Network Dataset Collection (<https://snap.stanford.edu/data/#socnets>)
2. Use the “Social Network: MOOC User Action Dataset” (<https://snap.stanford.edu/data/act-mooc.html>)
3. Load data in Neo4j – you can use Python/Java as the interface: reading data from files and send Cypher CREATE queries to Neo4j
4. Write and benchmark queries in Cypher for the following:
 - Show a small portion of your graph database (screenshot)
 - Count all users, count all targets, count all actions
 - Show all actions (actionID) and targets (targetID) of a specific user (choose one)
 - For each user, count his/her actions
 - For each target, count how many users have done this target
 - Count the average actions per user
 - Show the userID and the targetID, if the action has positive Feature2
 - For each targetID, count the actions with label “1”