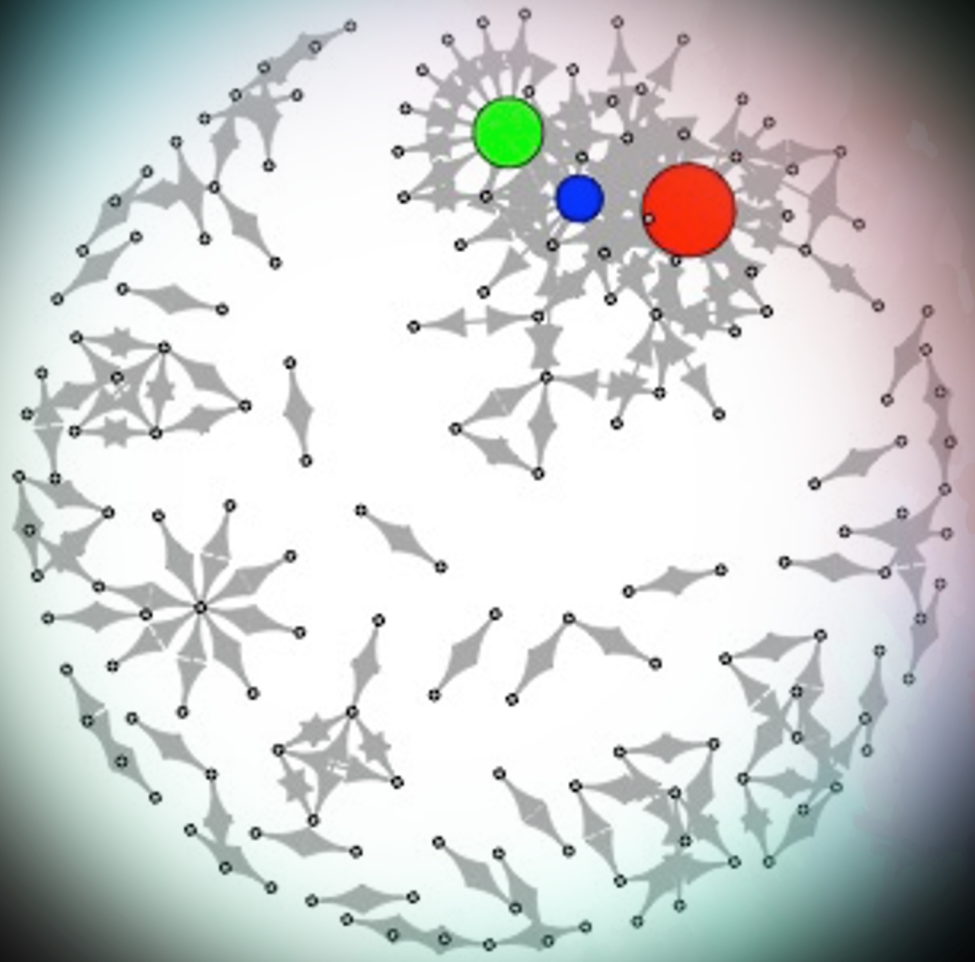
Centrality Analysis of IMDB Dataset



FUNSO OJE & OLUWAFEMI AJEIGBE

IMDB Dataset

7 datasets in repo

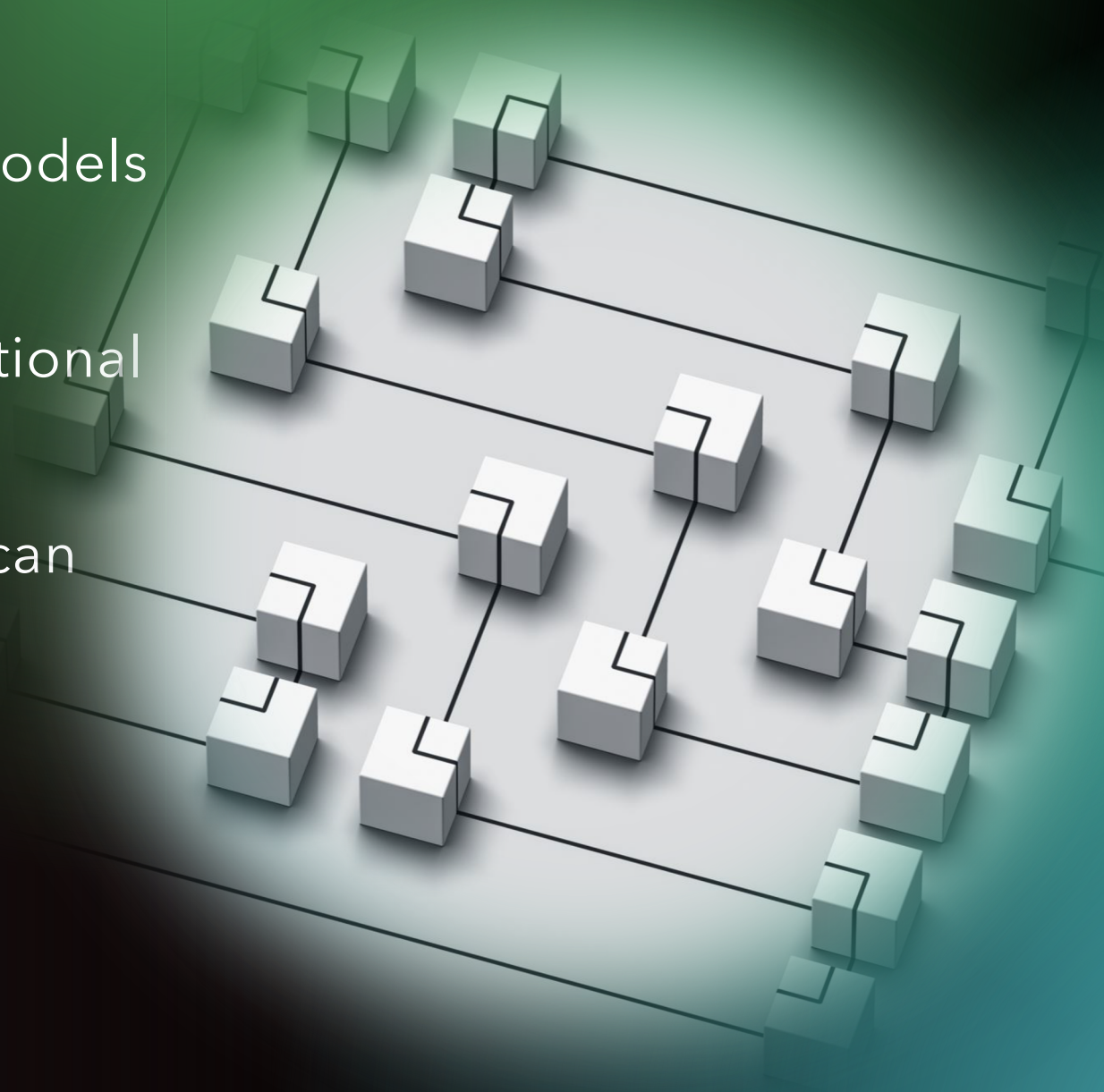
Dataset contains various title info e.g., rating, actors, crew, year, etc.

Filtered actors and actresses only (663k+ people)

Filtered movies only (600k+ titles)

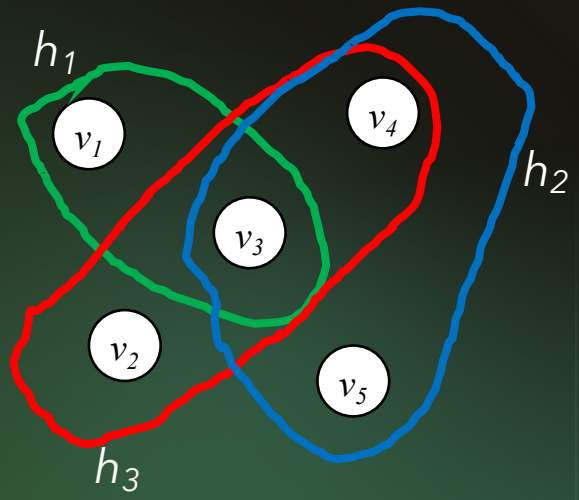
Motivation

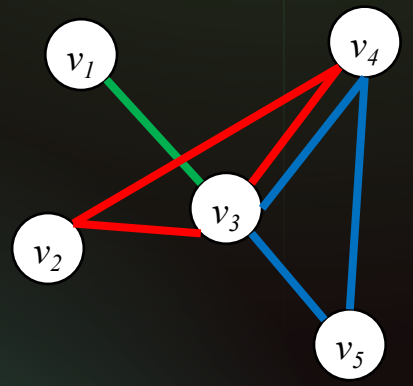
- Dataset Size
- Using Hypergraph – Simple models for Complex Data
- Modelling Complexity in Relational Data
- Using Parallel processing we can achieve faster runtimes

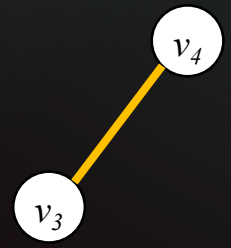


Data Analysis

- As a Hypergraph
- Adjacency Edge
- s-line graph (s=2)

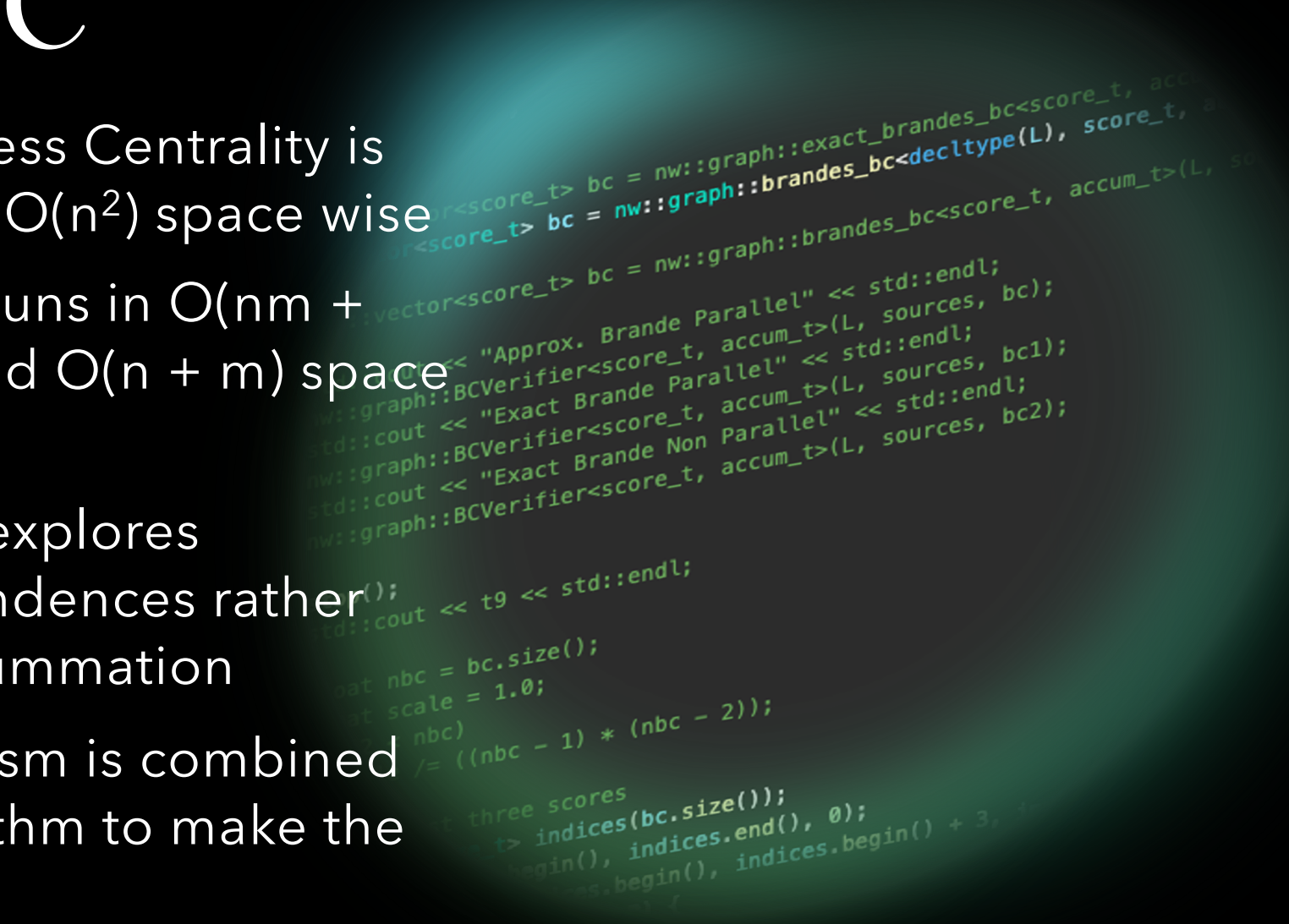






Brandes BC

- Typically, Betweenness Centrality is $O(n^3)$ time wise and $O(n^2)$ space wise
- Brandes Algorithm runs in $O(nm + n^2 log n)$ time wise and $O(n + m)$ space wise
- Brandes Algorithm explores accumulating dependences rather than dependency summation
- In addition, parallelism is combined with Brandes Algorithm to make the process quicker

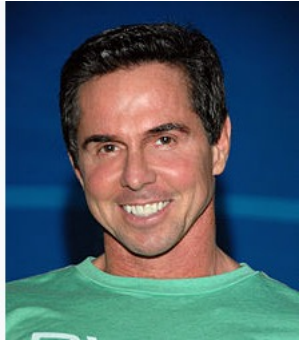


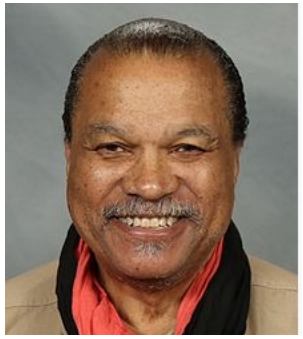
Results

Challenges

- Big Dataset
- s=1 issue
- Attempted to use a smaller dataset by filtering by movies made in the US

The US IMDB Dataset





What's next?

01

Validate the result

02

The true centrality result will be when s=1

03

Does the s-line approximation help to provide faster centrality results?

04

Where do the actors in a particular s-line fall when s=1?

Thank You!

- Questions and Discussions
- Many thanks to Tony and the PNNL team for their support