Graph Clustering Toolkit for Biological Data Analysis

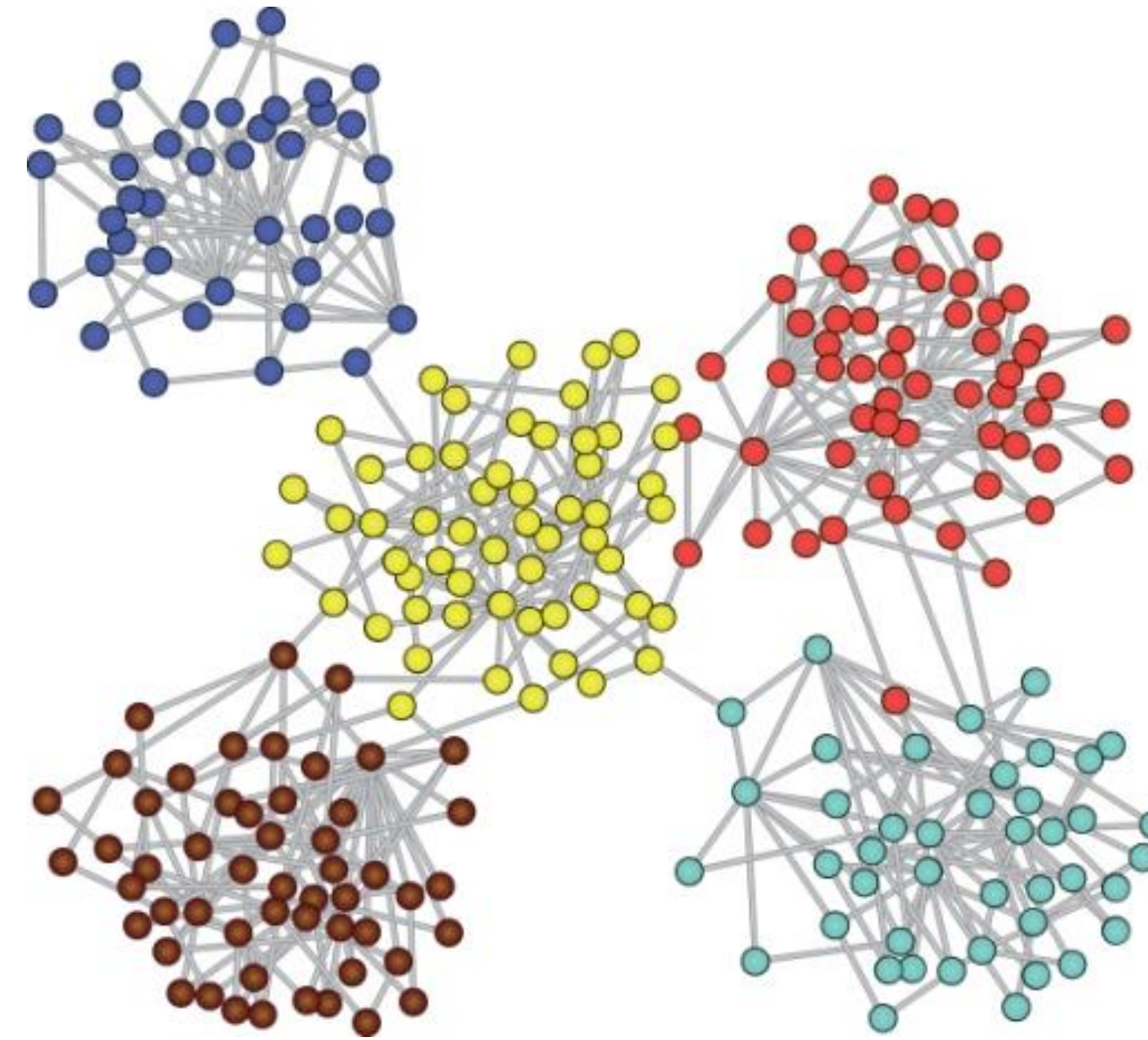Brian Rappaport, Anuththari Gamage

Tufts University ECE

Motivation

- Graph clustering helps us make important inferences about data sets

- We aggregate graph clustering algorithms to aid in machine learning research

- Our GUI provides an easy, intuitive way to study these algorithms

Protein Interaction Graphs

- Graphs represent relationships between different entities by vertices and edges

Transcription and RNA biosynthesis

Chromatin organization and modification

- GO: RNA Metabolism
- GO: Primary Metabolism
- GO: Protein Biosynthesis
- GO: DNA Metabolism
- GO: Transcription
- GO: Cell Cycle
- GO: Other
- GO: Biological Process
- GO: Signal Transduction
- GO: Transport

- Various types of interactions between proteins in a cell can be represented well using graphs

- Using these graphs, we can classify proteins based on their functions
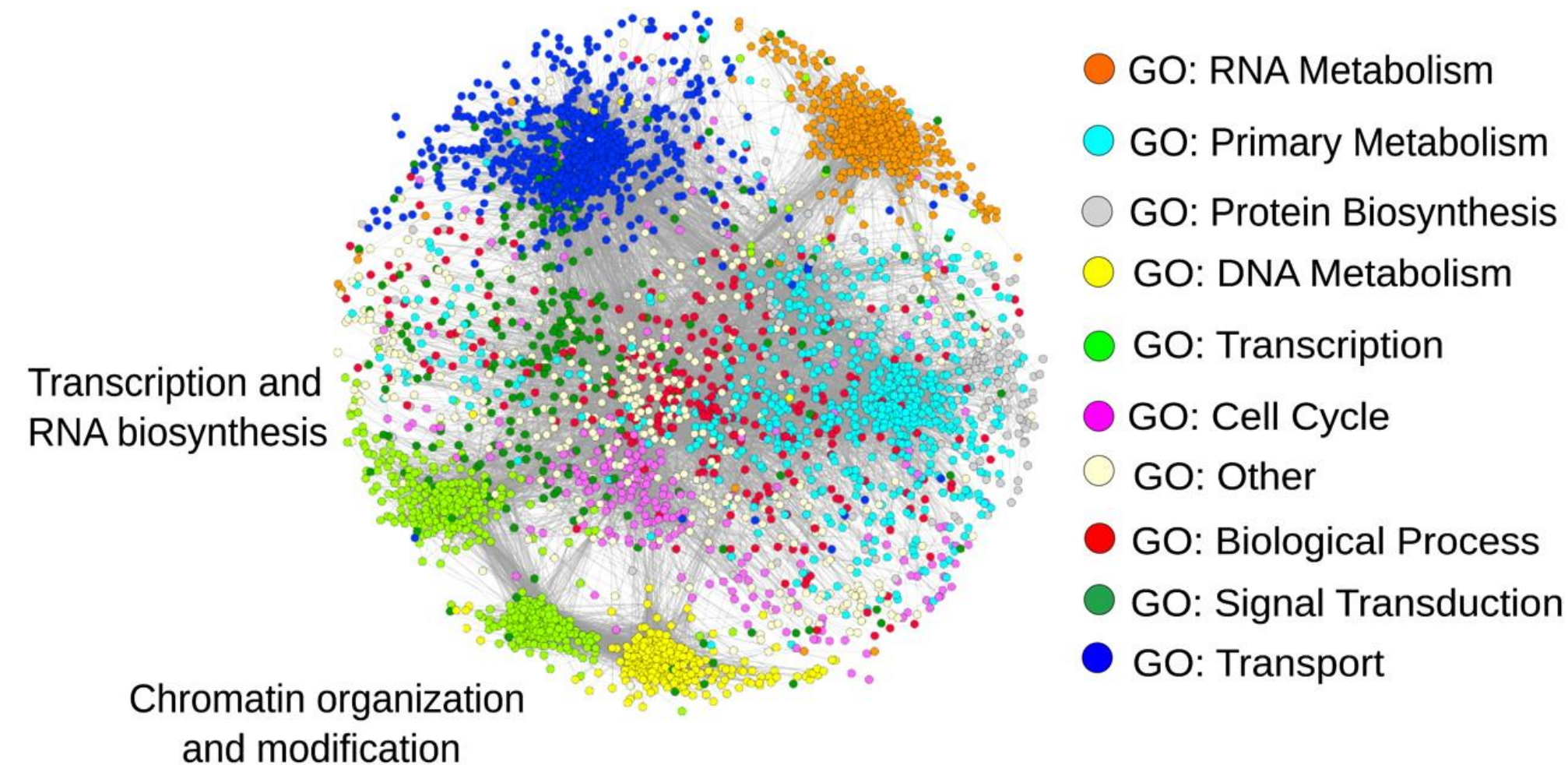
Graph Clustering

- We can group vertices into different groups, known as clusters, based on edge connectivity patterns.

- These clusters identify latent features like political affiliations in social networks and function categories in protein-protein interactions networks.

Graph Clustering Algorithms

- Based on matrix factorization
  - Spectral Clustering
  - Diffusion State Distance

- Based on random walks on graphs
  - Vec-NBT
  - Randomized Shortest Paths

Demo

Applications

- Ideal for graph clustering when the number of clusters in the data is known

- Can be used for data classification, link prediction, visualization

- Easy introduction to unsupervised machine learning and its uses

Future Work

- Extend algorithms to work with mixed-membership clusters (vertices belonging to multiple clusters)

- Integrate with multilayer-graph clustering algorithms, which improve prediction accuracy by combining multiple graphs