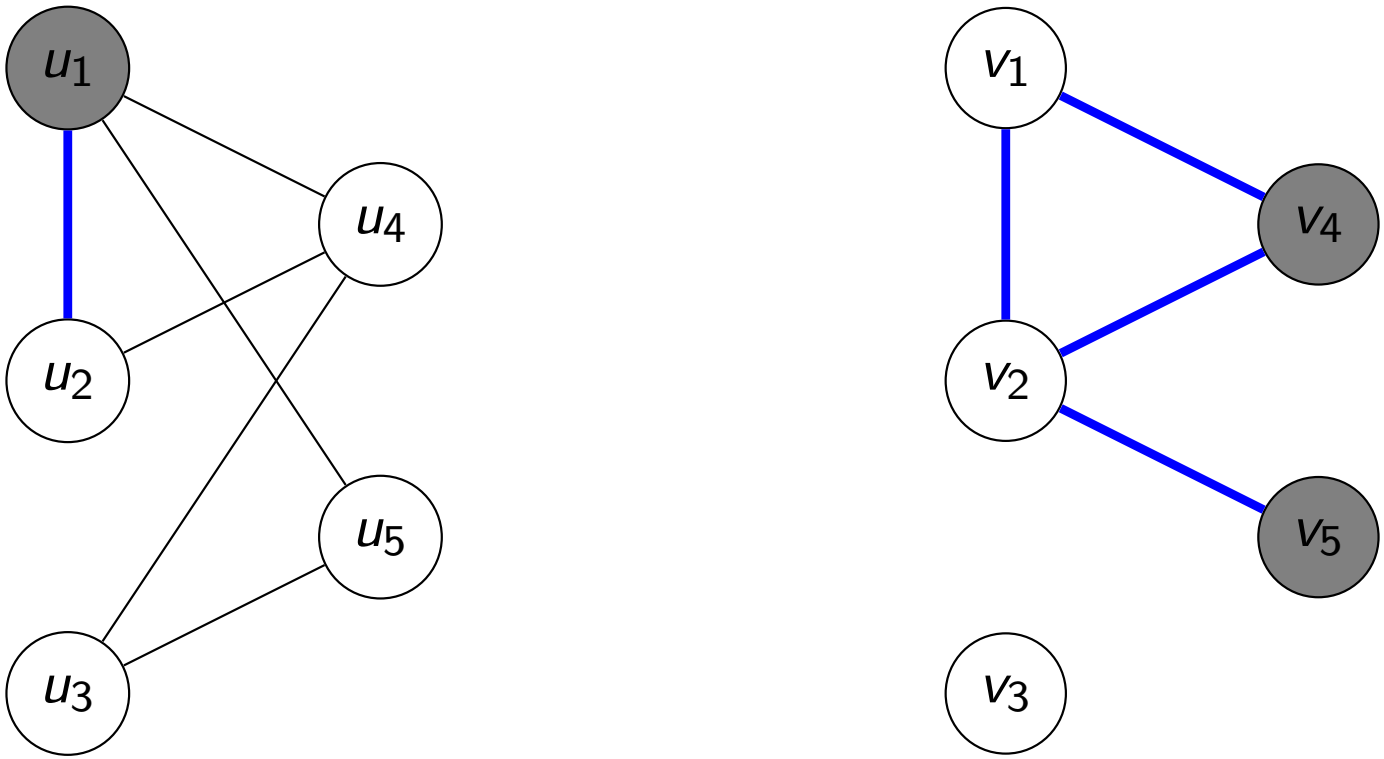
Algorithm Selection for Maximum Common Subgraph

Paulius Dilkas
Supervisors: Dr Patrick Prosser and Dr Ciaran McCreesh

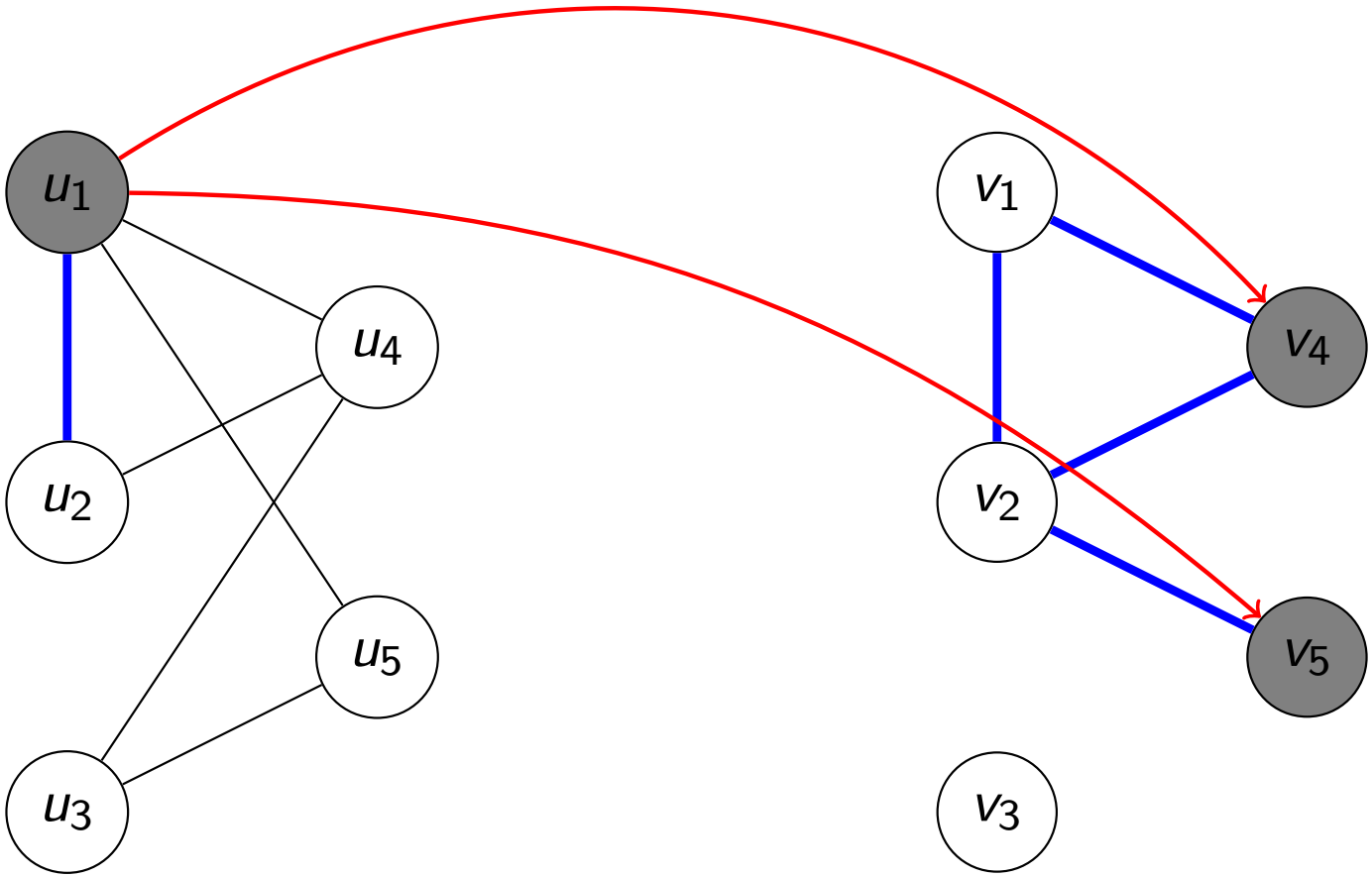
School of Computing Science
University of Glasgow

23rd March 2018

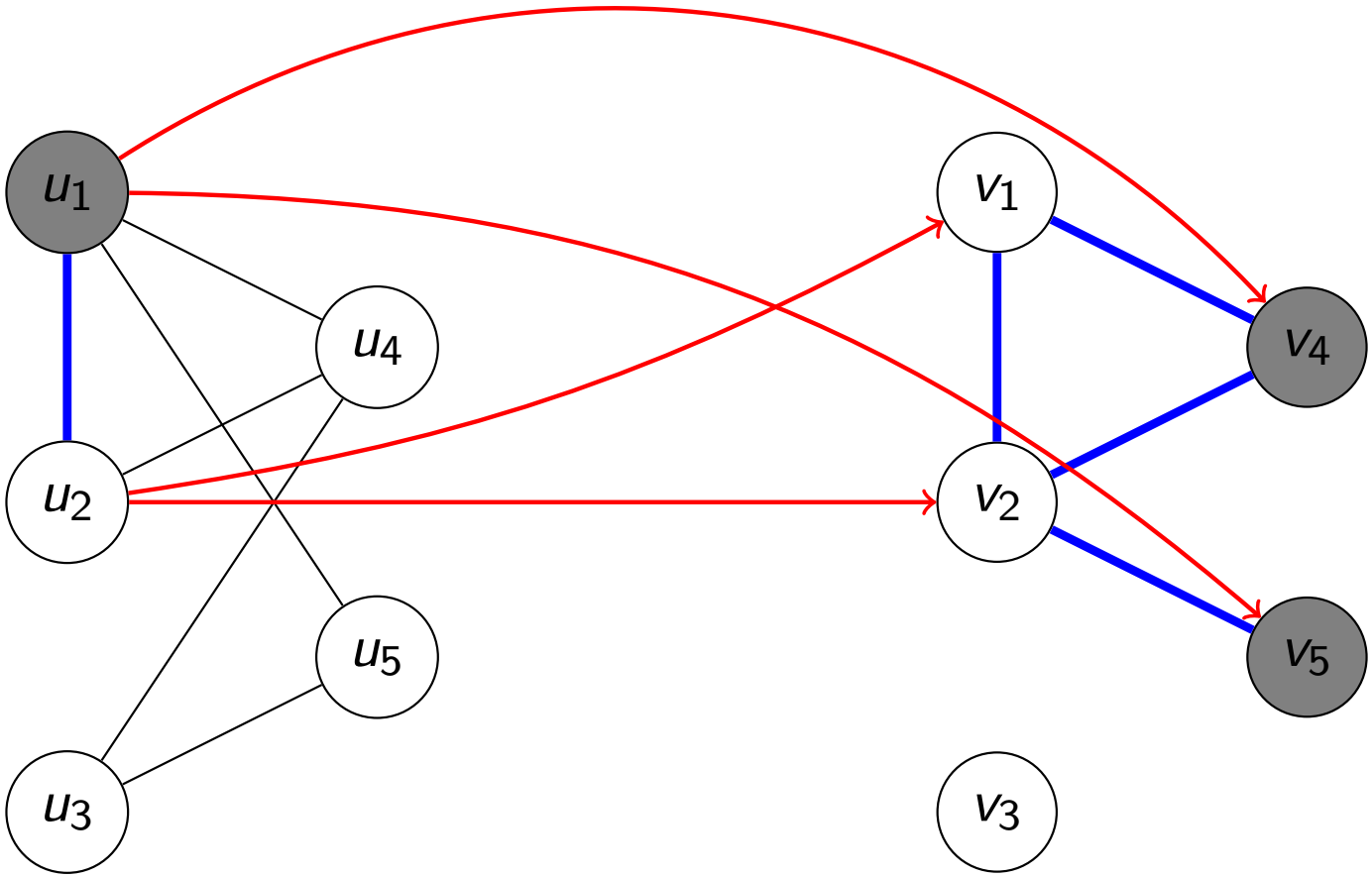
The Problem: Maximum Common Subgraph



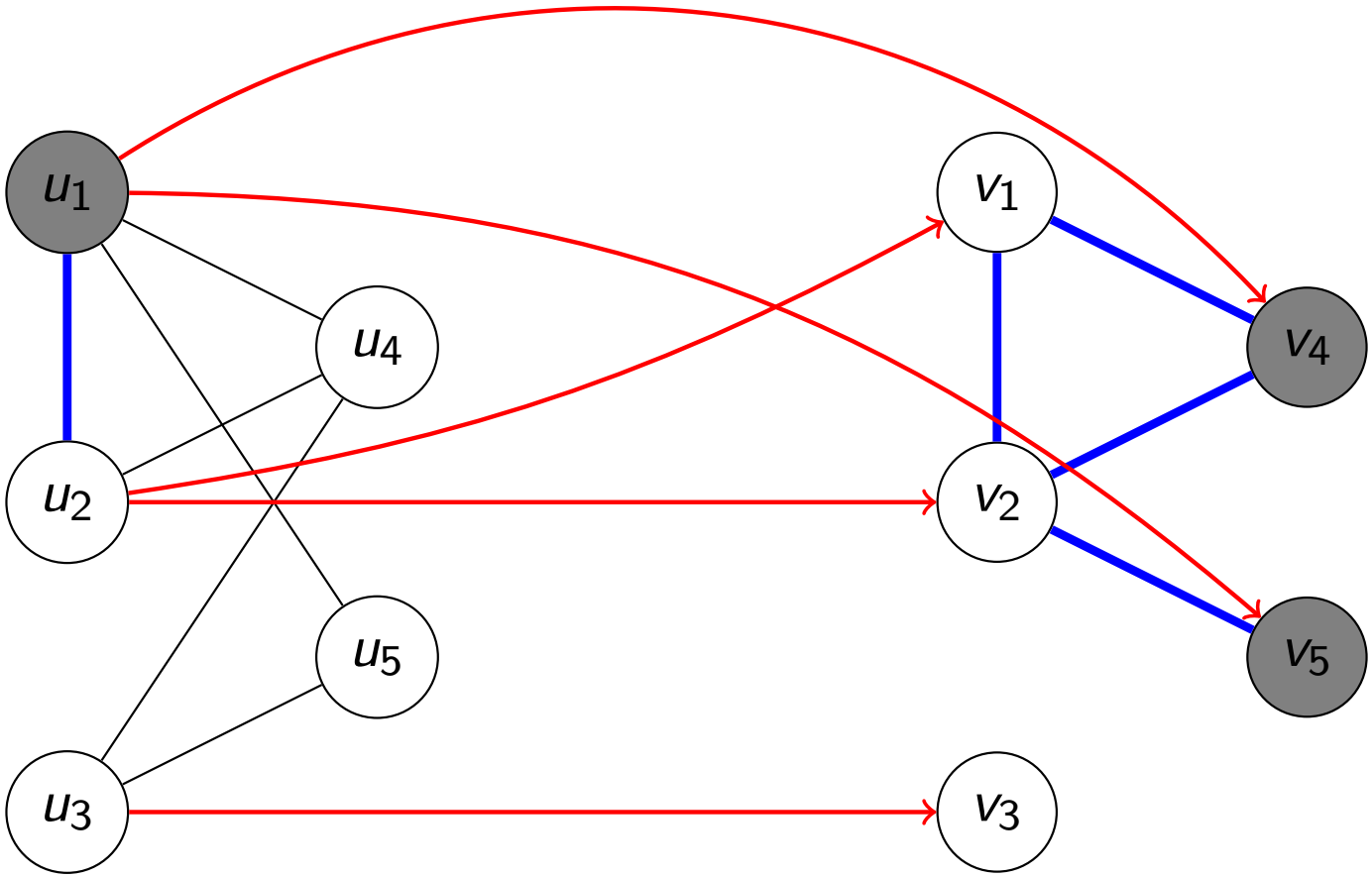
The Problem: Maximum Common Subgraph



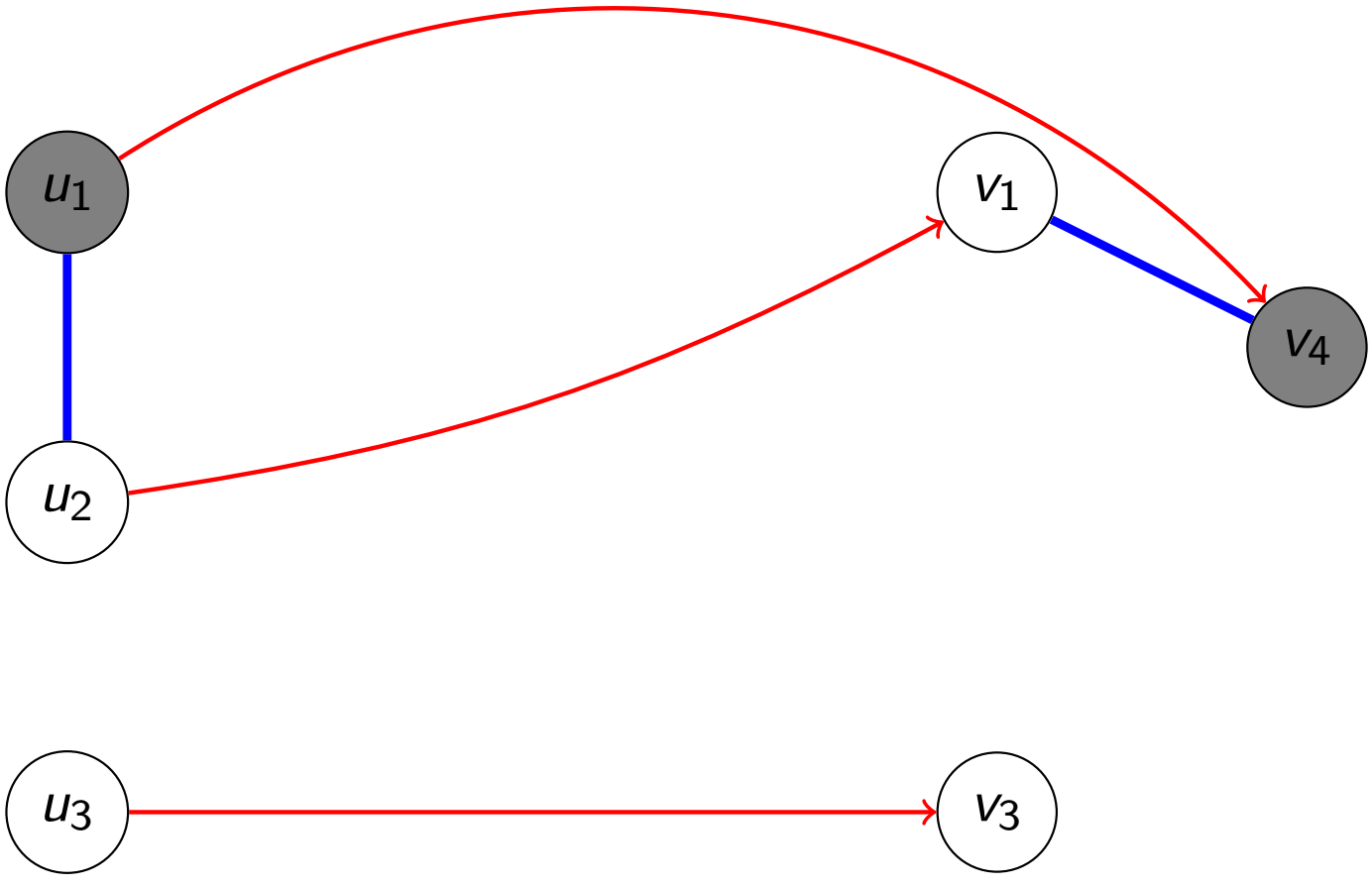
The Problem: Maximum Common Subgraph



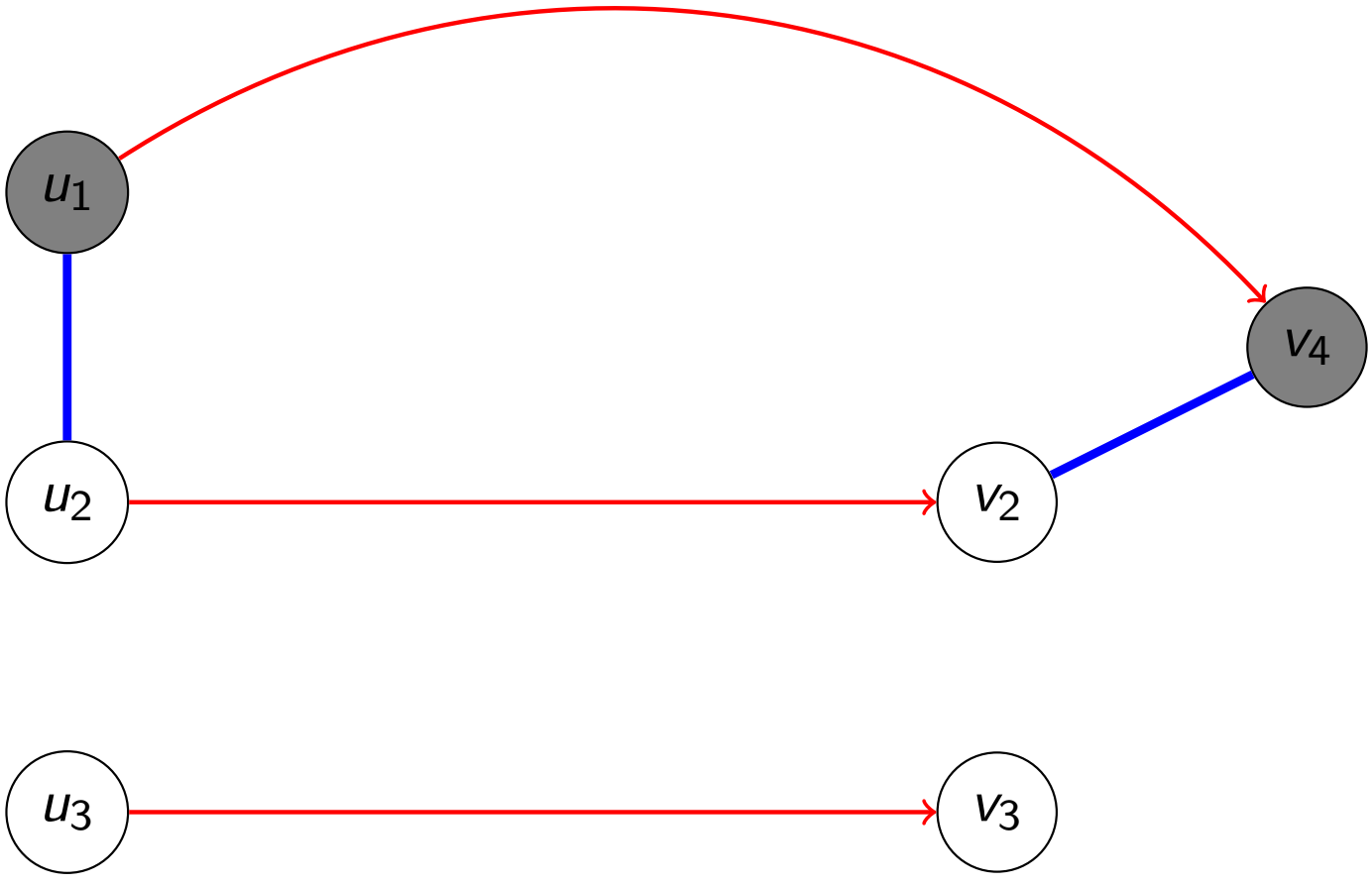
The Problem: Maximum Common Subgraph



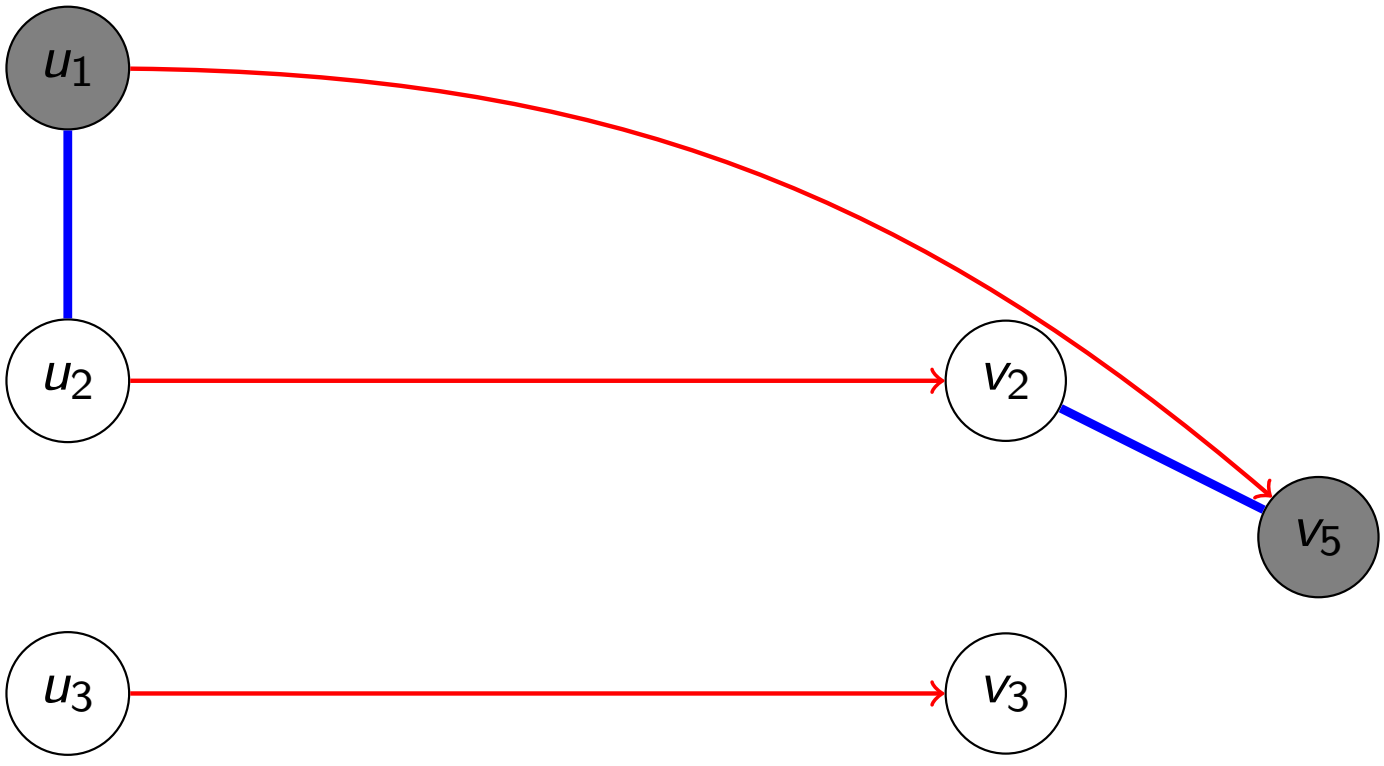
The Problem: Maximum Common Subgraph



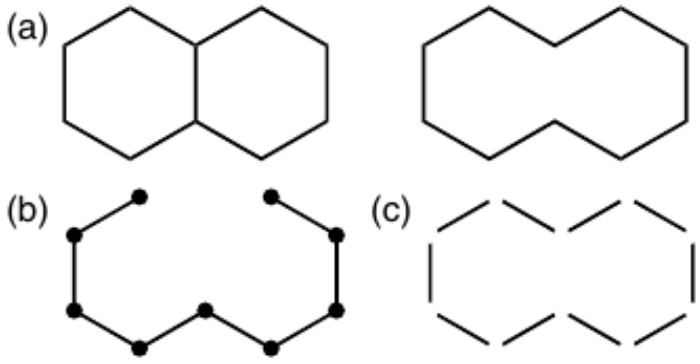
The Problem: Maximum Common Subgraph



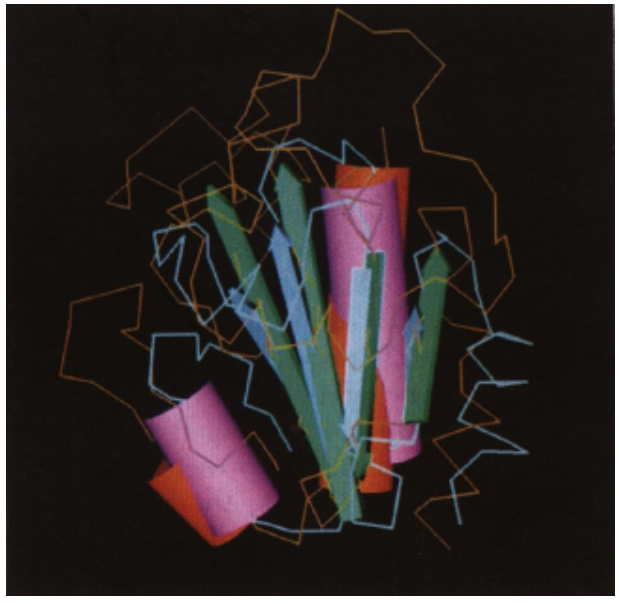
The Problem: Maximum Common Subgraph



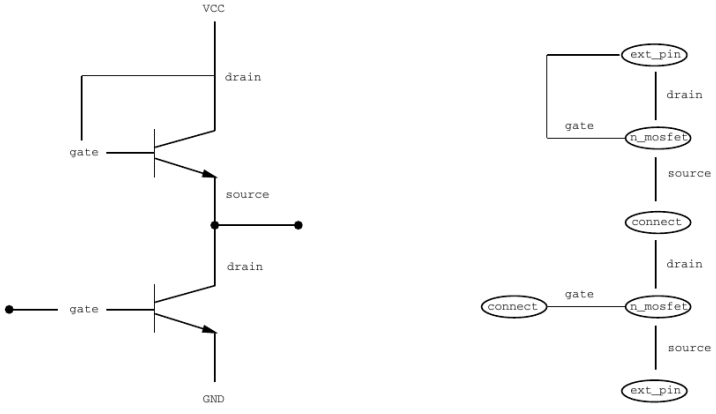
Why Is It Important?



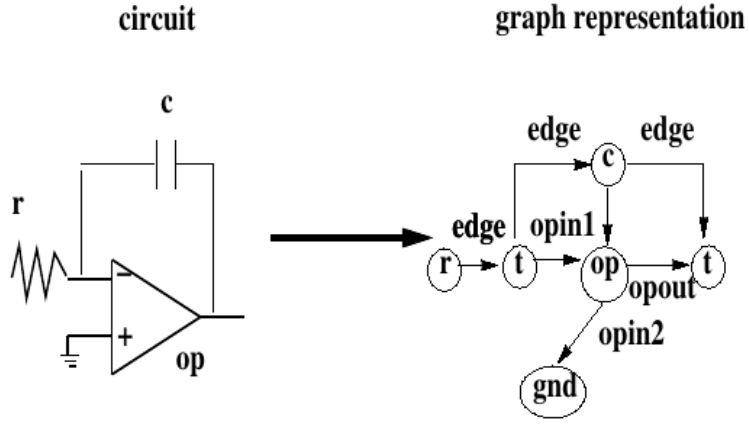
Source: Ehrlich and Rarey 2011



Source: M. Grindley et al. 1993



Source: Cook and Holder 1994



Source: Djoko, Cook and Holder 1997

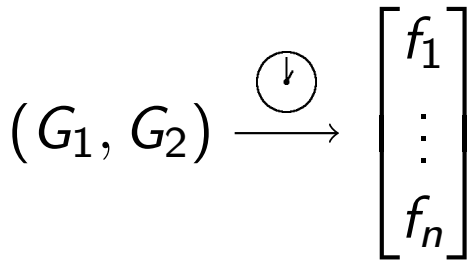
Existing Ways to Solve It

- McSplit, McSplit↓
  - McCreesh, Prosser and Trimble 2017
- clique encoding
  - McCreesh, Ndiaye et al. 2016
- k↓
  - Hoffmann, McCreesh and Reilly 2017

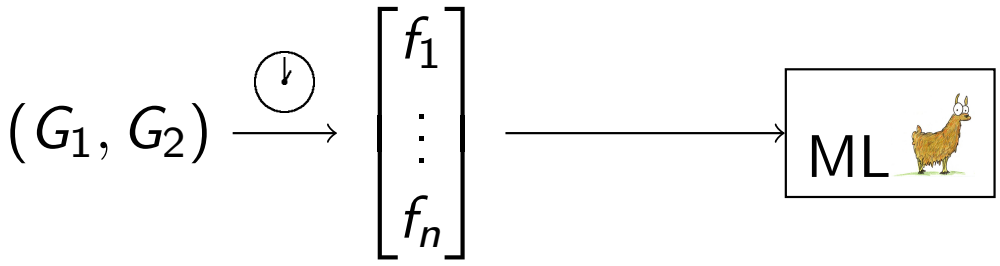
Solution: Algorithm Selection

(G_1, G_2)

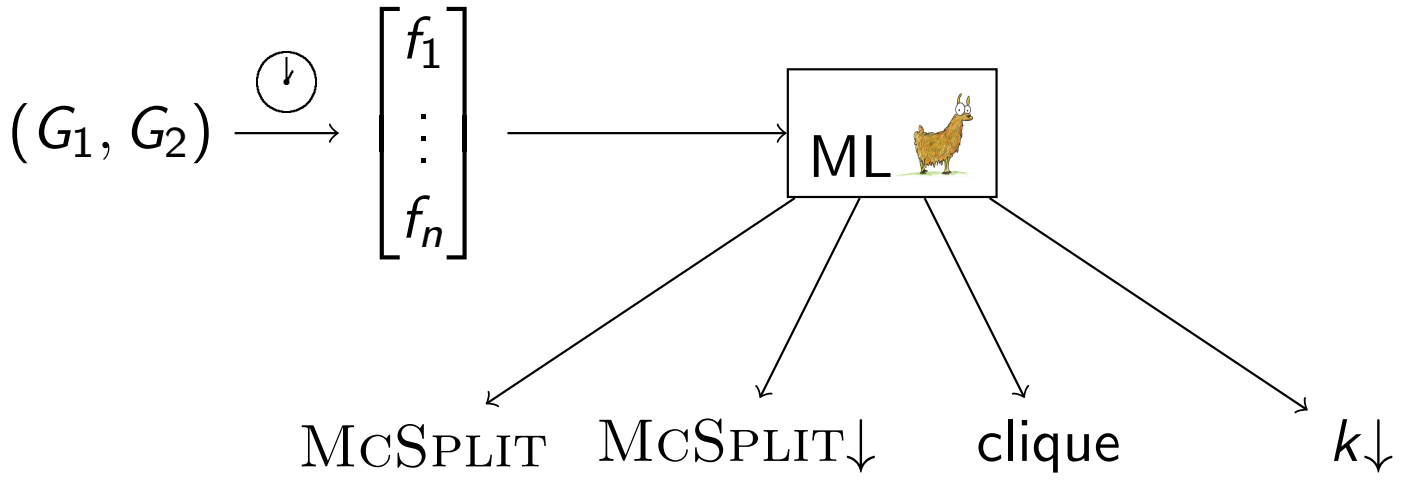
Solution: Algorithm Selection



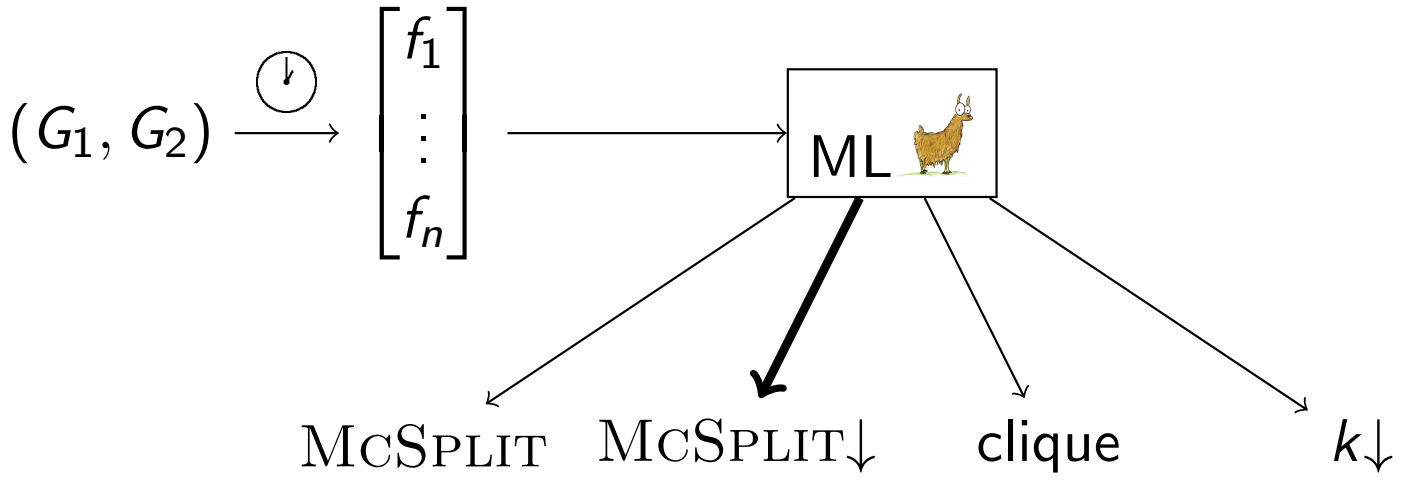
Solution: Algorithm Selection



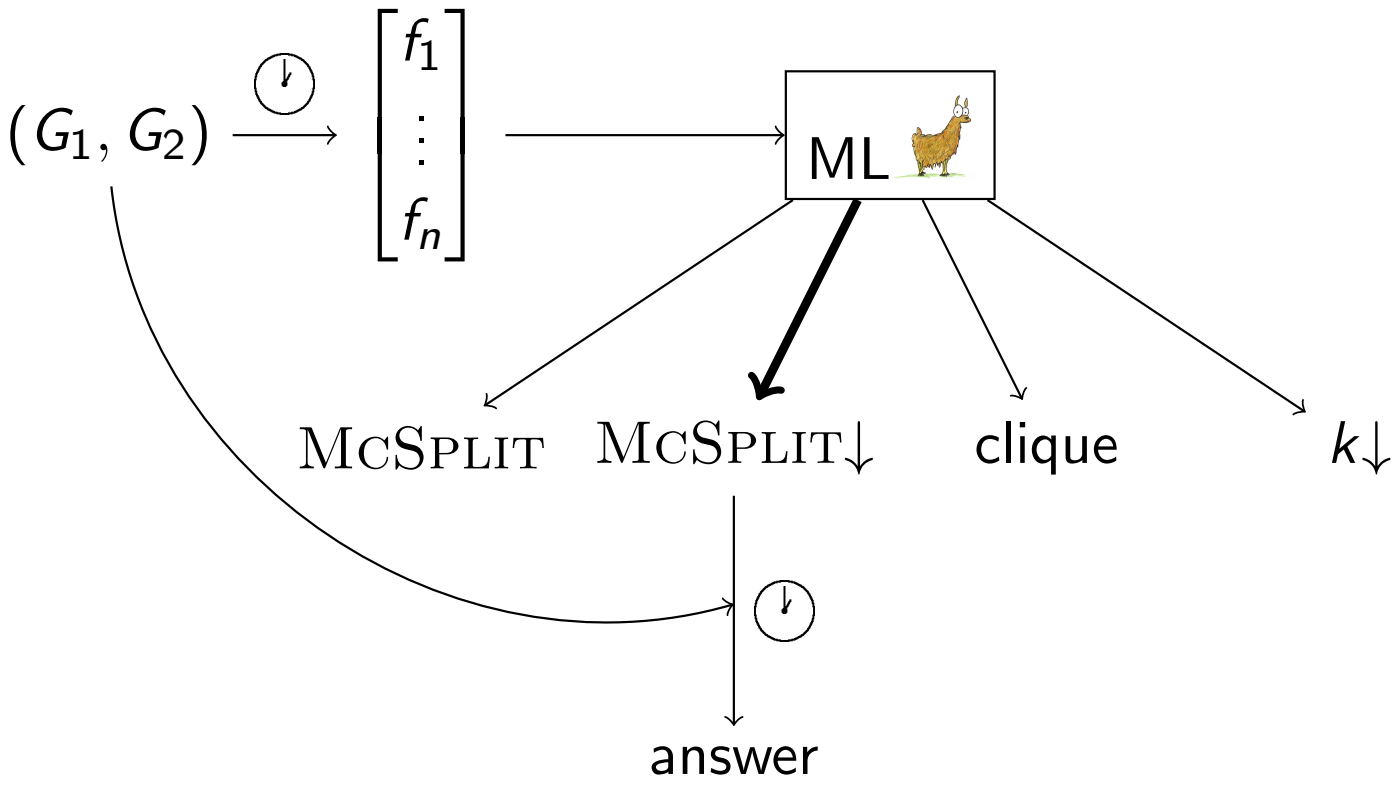
Solution: Algorithm Selection



Solution: Algorithm Selection



Solution: Algorithm Selection



The Process

- Three cases:
  - no labels
  - vertex labels
  - vertex and edge labels

The Process

- Three cases:
  - no labels
  - vertex labels
  - vertex and edge labels
- How many different labels?
  - Measured as a percentage of the number of vertices/edges
  - 5%, 10%, 15%, 20%, 25%, 33%, 50%

The Process

- Three cases:
  - no labels
  - vertex labels
  - vertex and edge labels
- How many different labels?
  - Measured as a percentage of the number of vertices/edges
  - 5%, 10%, 15%, 20%, 25%, 33%, 50%
- Run every algorithm on every instance
  - $\sim$ 500,000 experiments for each algorithm

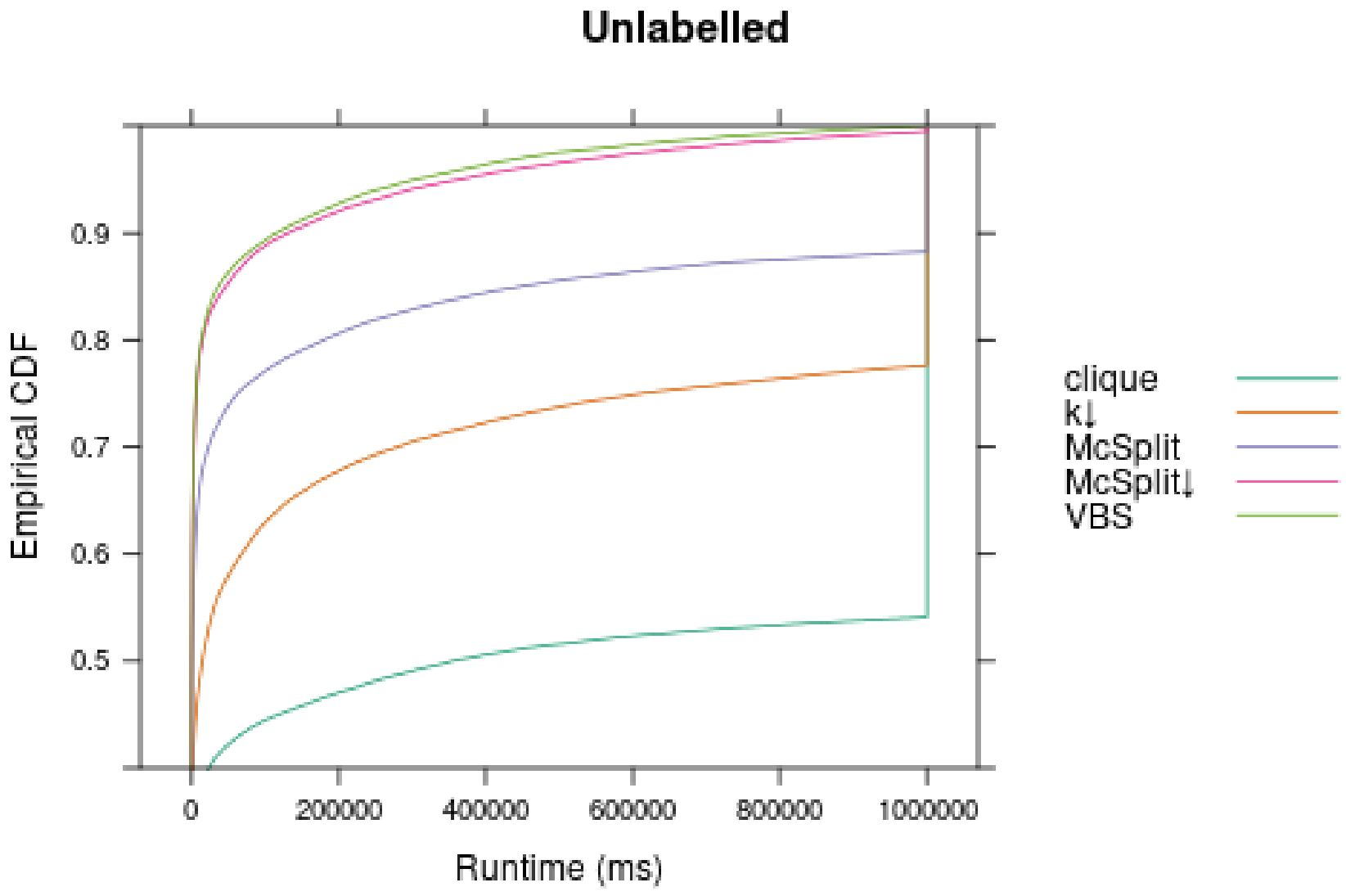
The Process

- Three cases:
  - no labels
  - vertex labels
  - vertex and edge labels
- How many different labels?
  - Measured as a percentage of the number of vertices/edges
  - 5%, 10%, 15%, 20%, 25%, 33%, 50%
- Run every algorithm on every instance
  - $\sim$ 500,000 experiments for each algorithm
- Identify and extract features
  - 34 in total
  - some from Kotthoff, McCreesh and Solnon 2016
  - some new

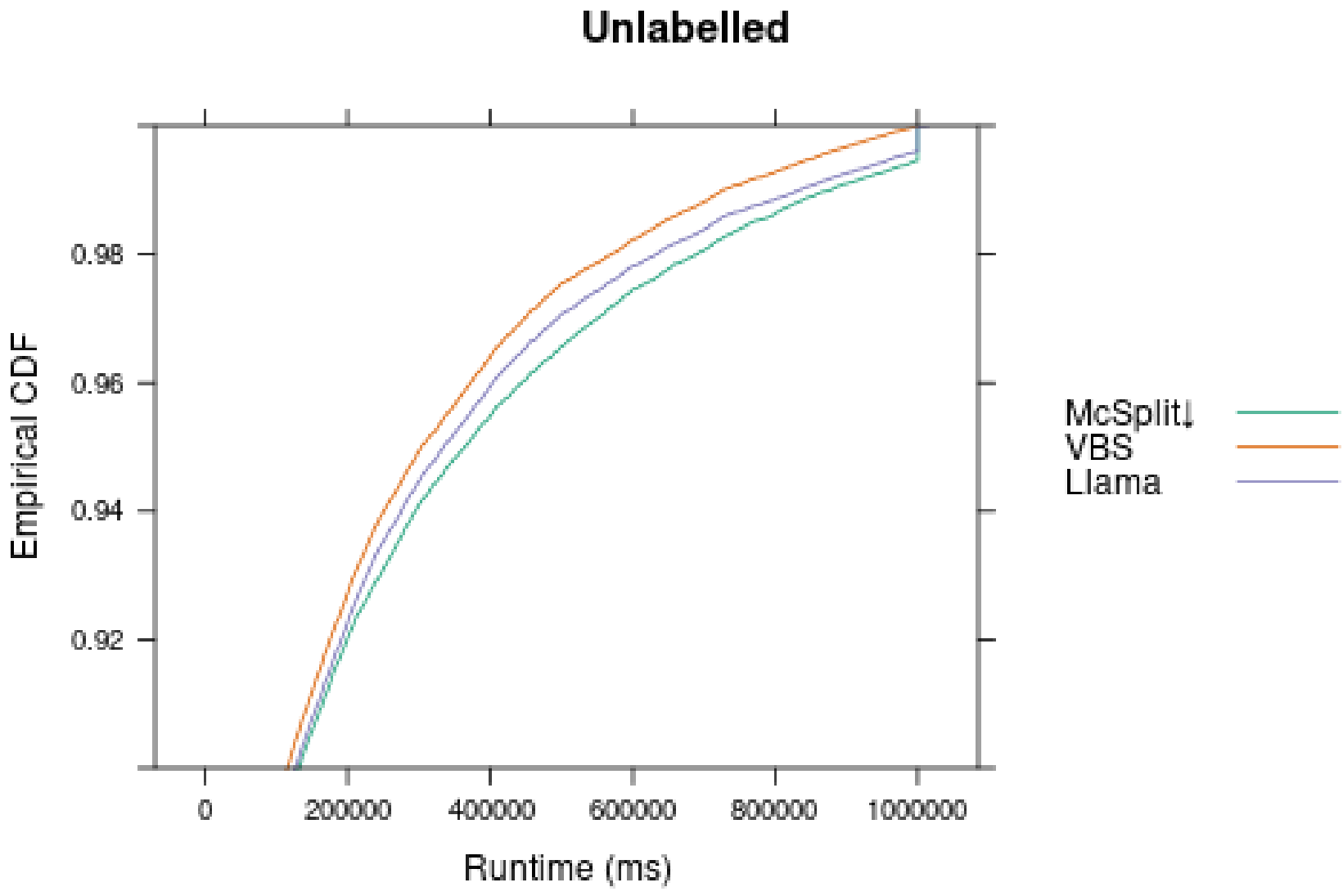
The Process

- Three cases:
  - no labels
  - vertex labels
  - vertex and edge labels
- How many different labels?
  - Measured as a percentage of the number of vertices/edges
  - 5%, 10%, 15%, 20%, 25%, 33%, 50%
- Run every algorithm on every instance
  - $\sim$ 500,000 experiments for each algorithm
- Identify and extract features
  - 34 in total
  - some from Kotthoff, McCreesh and Solnon 2016
  - some new
- Train machine learning models
- Evaluate their performance and usefulness

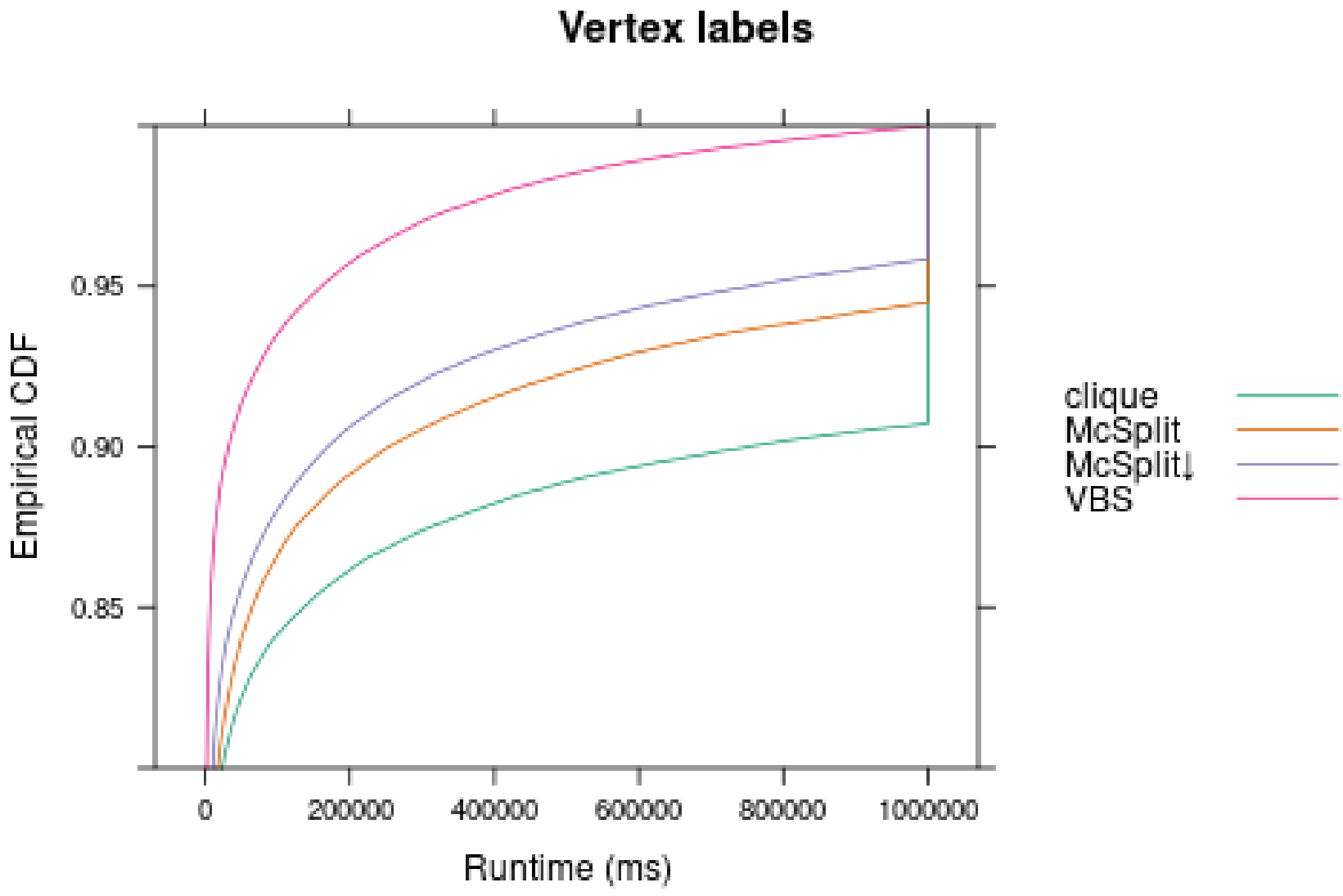
Results


Unlabelled
Empirical CDF
0.9
0.8
0.7
0.6
0.5
0
200000
400000
600000
800000
1000000
Runtime (ms)
clique
k↓
McSplit
McSplit↓
VBS

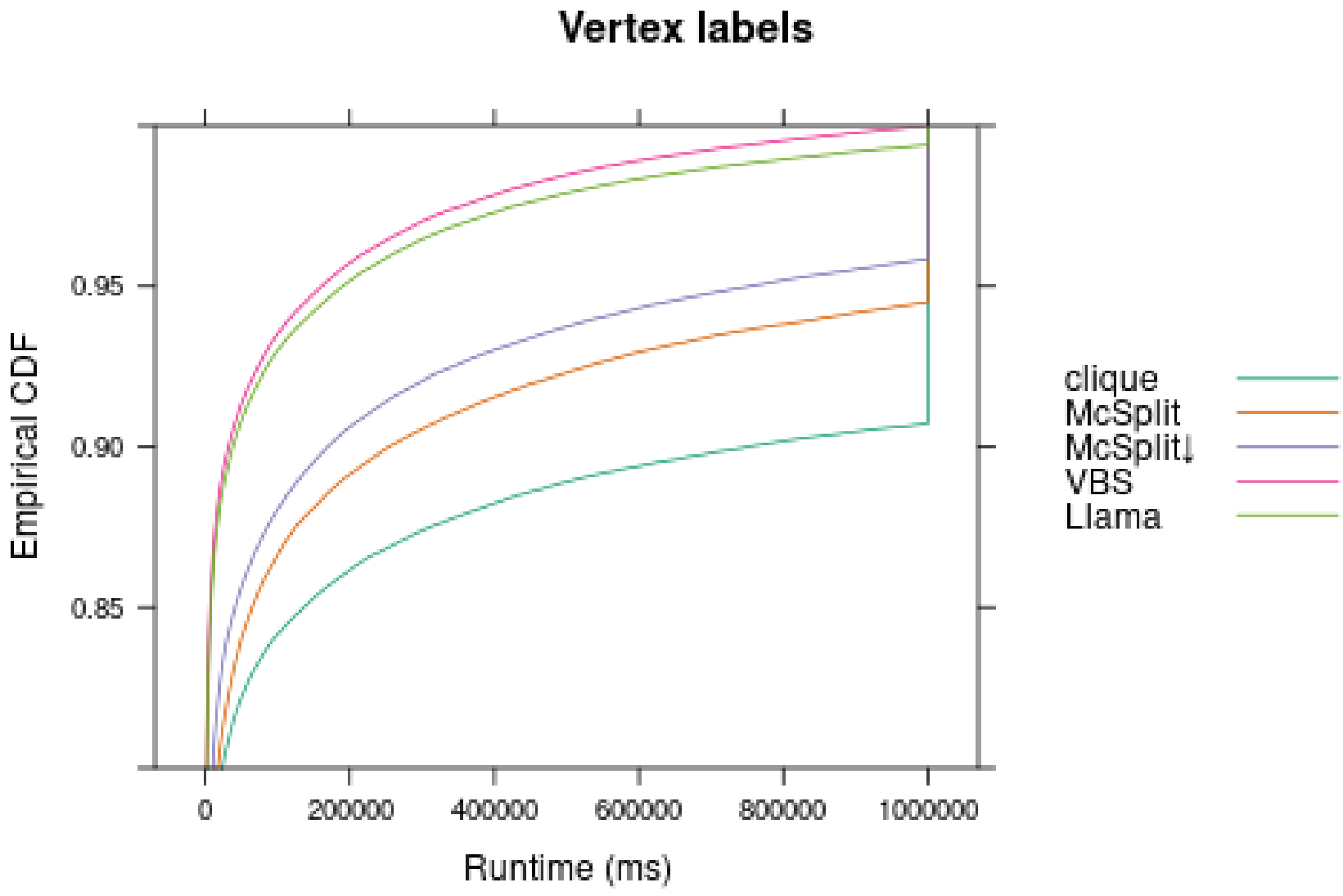
Results (27%)


Unlabelled
Empirical CDF
0.98
0.96
0.94
0.92
0
200000
400000
600000
800000
1000000
Runtime (ms)
McSplit↓
VBS
Llama

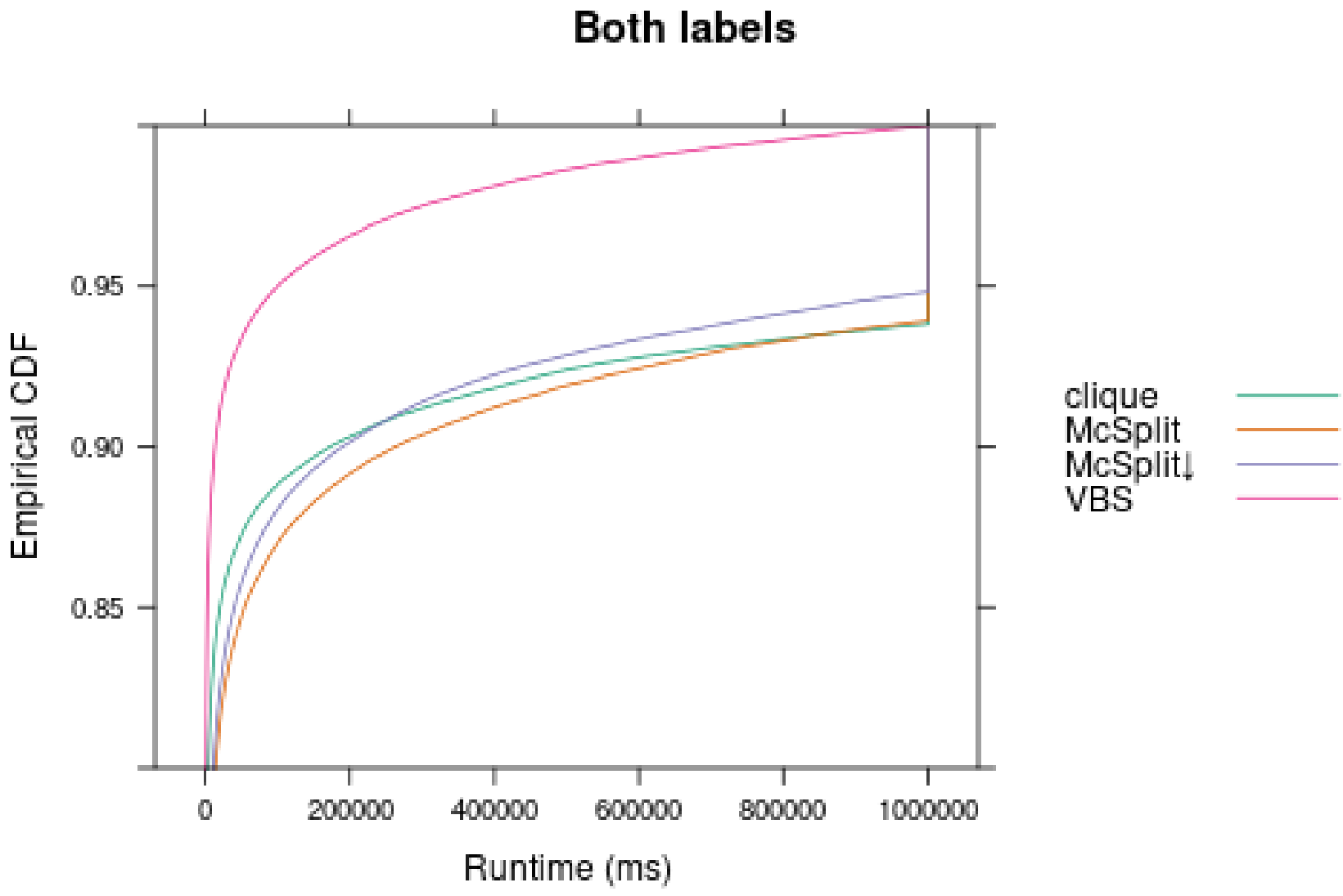
Results


Vertex labels
Empirical CDF
0.95
0.90
0.85
0
200000
400000
600000
800000
1000000
Runtime (ms)
clique
McSplit
McSplit↓
VBS

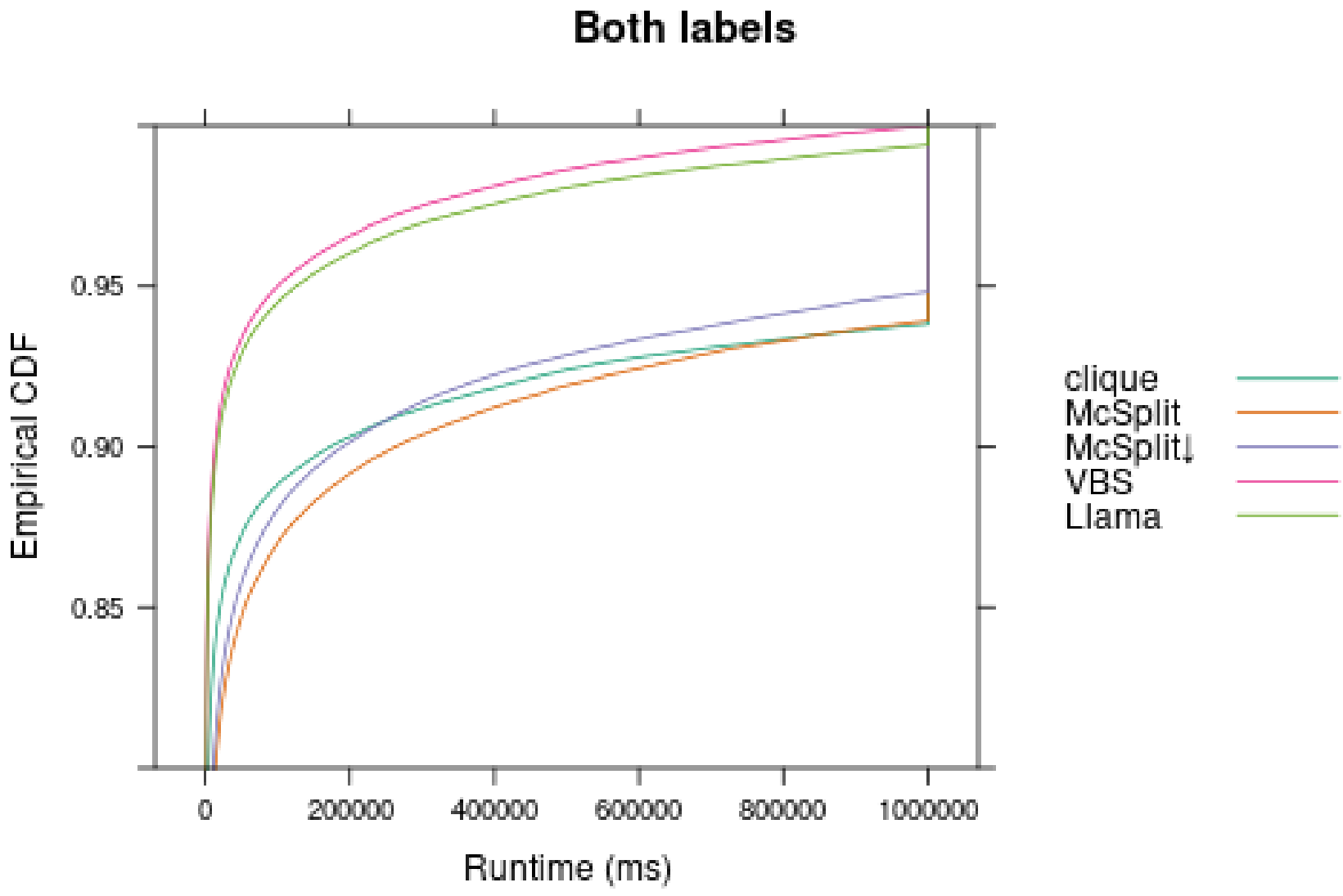
Results (86%)


Vertex labels
Empirical CDF
0.95
0.90
0.85
0
200000
400000
600000
800000
1000000
Runtime (ms)
clique
McSplit
McSplit↓
VBS
Llama

Results


Both labels
Empirical CDF
0.95
0.90
0.85
0
200000
400000
600000
800000
1000000
Runtime (ms)
clique
McSplit
McSplit↓
VBS

Results (88%)


Both labels
Empirical CDF
0.95
0.90
0.85
0
200000
400000
600000
800000
1000000
Runtime (ms)
clique
McSplit
McSplit↓
VBS
Llama

Other Accomplishments

- Identified important features:
  - labelling percentage, standard deviation of degrees
- Discovered how algorithms' performance changes with fewer labels
- Extended $k\downarrow$ to support vertex labels
  - using neighbourhood degree sequences
- Defined and developed new algorithms capable of switching between McSplit and the clique encoding