Algorithm Selection for Maximum Common Subgraph

Paulius Dilkas

FATA seminar

16th January 2018

Algorithm selection

Definition (Bischl et al. 2016)

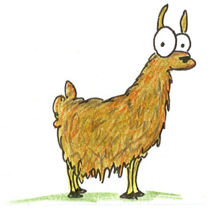
Given a set $\mathcal{I}$ of problem instances, a space of algorithms $\mathcal{A}$, and a performance measure $m\colon \mathcal{I} \times \mathcal{A} \to \mathbb{R}$, the *algorithm selection problem* is to find a mapping $s\colon \mathcal{I} \to \mathcal{A}$ that optimises $\mathbb{E}[m(i, s(i))]$.

Algorithm selection

Definition (Bischl et al. 2016)

Given a set $\mathcal{I}$ of problem instances, a space of algorithms $\mathcal{A}$, and a performance measure $m\colon \mathcal{I} \times \mathcal{A} \to \mathbb{R}$, the *algorithm selection problem* is to find a mapping $s\colon \mathcal{I} \to \mathcal{A}$ that optimises $\mathbb{E}[m(i, s(i))]$.

LLAMA (Kotthoff 2013)



Algorithms

- McSplit, McSplit $\downarrow$
  - (McCreesh, Prosser and Trimble 2017)
- clique encoding
  - (McCreesh, Ndiaye et al. 2016)
- $k\downarrow$
  - (Hoffmann, McCreesh and Reilly 2017)

Labelling

Data from Foggia, Sansone and Vento 2001; Santo et al. 2003 (81,400 pairs of graphs)

Labelling

Data from Foggia, Sansone and Vento 2001; Santo et al. 2003 (81,400 pairs of graphs)

Definition

A *vertex-labelled graph* is a 3-tuple $G = (V, E, \mu)$, where $\mu \colon V \to \{0, \dots, N-1\}$ is a vertex labelling function, for some $N \in \{1, \dots, |V|\}$.

Labelling

Data from Foggia, Sansone and Vento 2001; Santo et al. 2003 (81,400 pairs of graphs)

Definition

A *vertex-labelled graph* is a 3-tuple $G = (V, E, \mu)$, where $\mu \colon V \to \{0, \dots, N-1\}$ is a vertex labelling function, for some $N \in \{1, \dots, |V|\}$.

Definition

A graph $G = (V, E, \mu)$ is said to have a *p% (vertex) labelling* if

$$N = \max\left\{2^n : n \in \mathbb{N},\, 2^n < \left\lfloor \frac{p}{100\%} \times |V| \right\rfloor\right\}.$$

Labelling

Definition

A graph $G = (V, E, \mu)$ is said to have a $p\%$ *(vertex) labelling* if

$$N = \max\left\{2^n : n \in \mathbb{N},\, 2^n < \left\lfloor \frac{p}{100\%} \times |V| \right\rfloor \right\}.$$

- 5% labelling - 20 vertices per label on average
- 50% labelling - 2 vertices per label on average

Labelling

Definition

A graph $G = (V, E, \mu)$ is said to have a *p% (vertex) labelling* if

$$N = \max \left\{ 2^n : n \in \mathbb{N},\, 2^n < \left\lfloor \frac{p}{100\%} \times |V| \right\rfloor \right\}.$$

- 5% labelling - 20 vertices per label on average
- 50% labelling - 2 vertices per label on average
- Typical values explored: 33%, 50%, 75%
    - Often just 33%

Labelling

Definition

A graph $G = (V, E, \mu)$ is said to have a $p\%$ *(vertex) labelling* if

$$N = \max\left\{2^n : n \in \mathbb{N},\ 2^n < \left\lfloor \frac{p}{100\%} \times |V| \right\rfloor\right\}.$$

- 5% labelling - 20 vertices per label on average
- 50% labelling - 2 vertices per label on average
- Typical values explored: 33%, 50%, 75%
  - Often just 33%
- In my data: 5%, 10%, 15%, 20%, 25%, 33%, 50%

Labelling

Definition

A graph $G = (V, E, \mu)$ is said to have a *p% (vertex) labelling* if

$$N = \max \left\{ 2^n : n \in \mathbb{N},\ 2^n < \left\lfloor \frac{p}{100\%} \times |V| \right\rfloor \right\}.$$

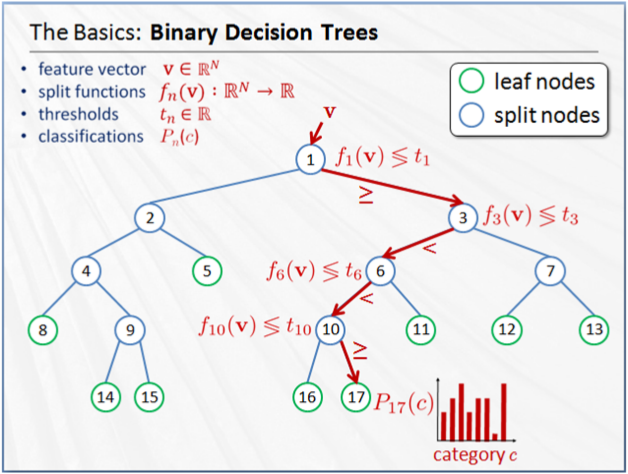
- 5% labelling - 20 vertices per label on average
- 50% labelling - 2 vertices per label on average
- Typical values explored: 33%, 50%, 75%
  - Often just 33%
- In my data: 5%, 10%, 15%, 20%, 25%, 33%, 50%
- 3 subproblems
  - no labels
  - vertex labels
  - vertex and edge labels

Features (34 in total)

The first 9 are from Kotthoff, McCreesh and Solnon 2016

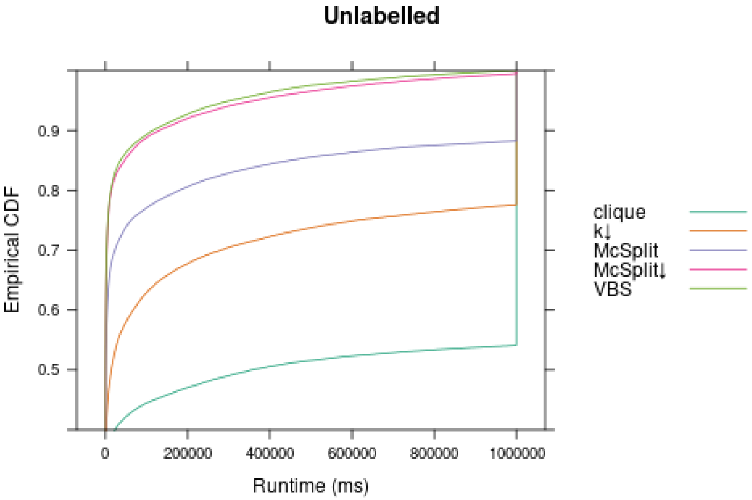
1. number of vertices
2. number of edges
3. mean/max degree
4. density
5. mean/max distance between pairs of vertices
6. standard deviation of degrees
7. number of loops
8. proportion of vertex pairs with distance $\geq 2, 3, 4$
9. connectedness
10. labelling percentage
11. ratios of features 1–5

Random forests

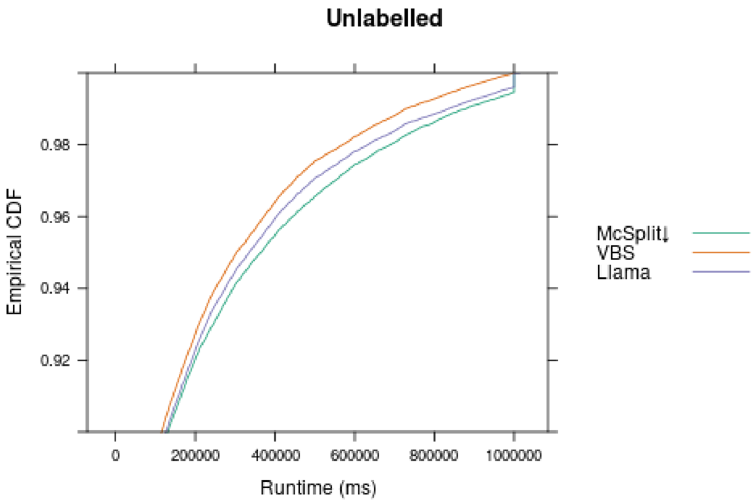


Source: Tae-Kyun Kim & Bjorn Stenger, Intelligent Systems and Networks (ISN) Research Group, Imperial College London

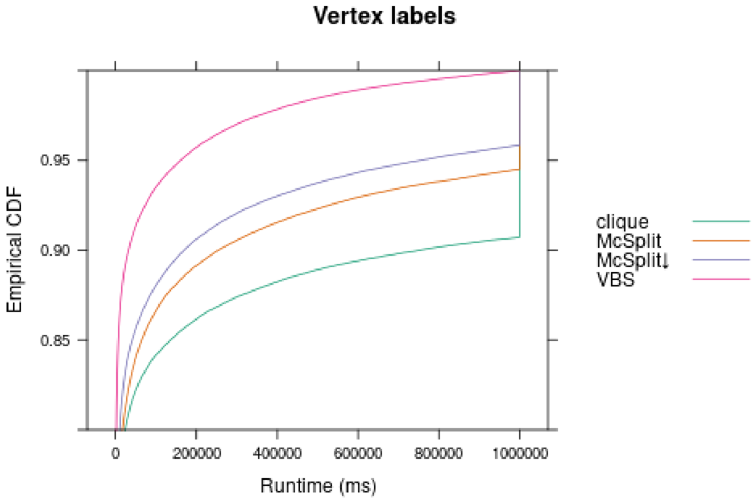
Results


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

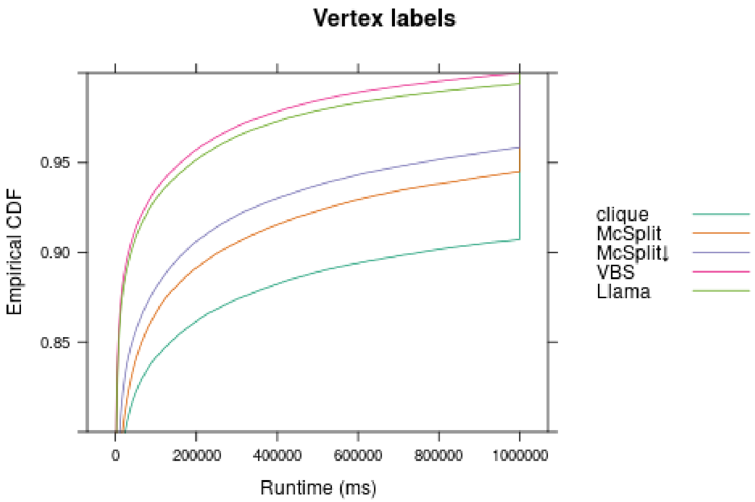
Results (27%)


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

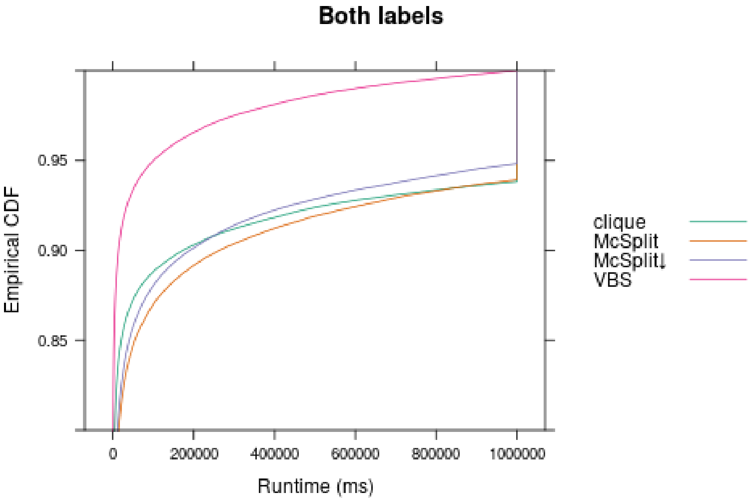
Results


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

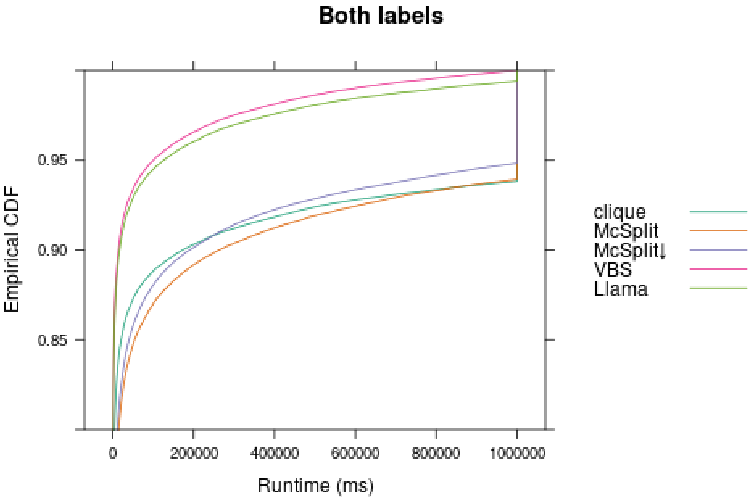
Results (86%)


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

Results


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

Results (88%)


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

Errors

- Out-of-bag
- (for each algorithm) $1 - \text{recall}$

Definition

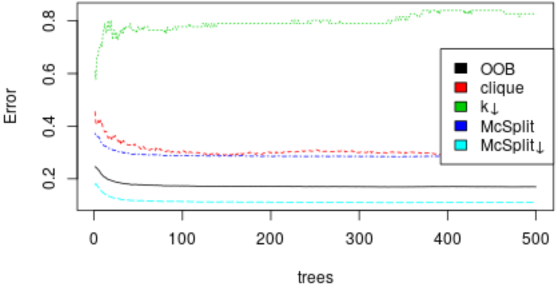
For an algorithm A, *recall* is

$$\frac{\text{the number of instances that were correctly predicted as } A}{\text{the number of instances where } A \text{ is the correct prediction}}.$$

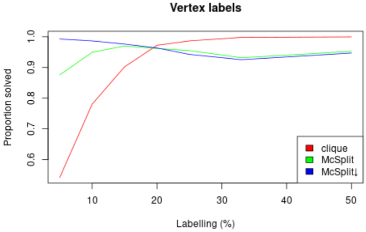
Errors (%)

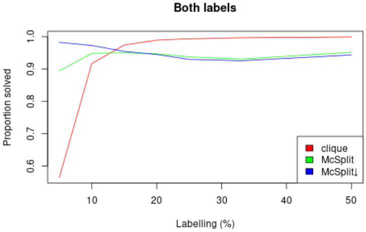
Error	Labelling		
	no	vertex	both
out-of-bag	17	13	14
clique	30	8	7
McSplit	29	22	29
McSplit ↓	11	11	11
k ↓	80		

Convergence of errors for unlabelled graphs

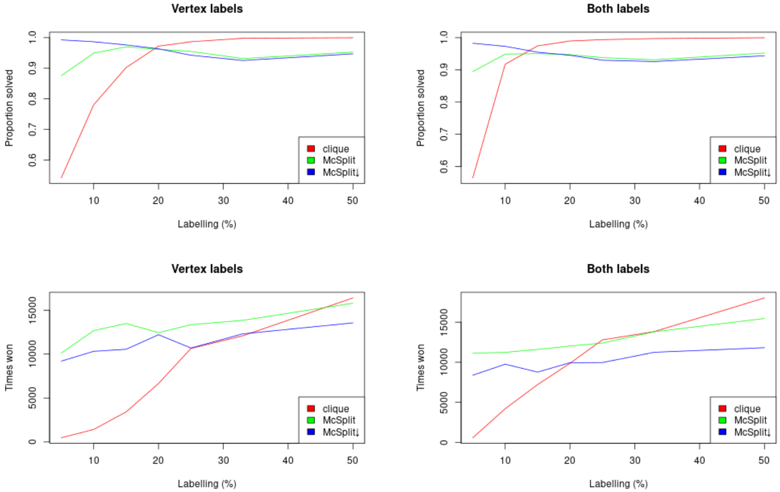


What happens when labelling changes?





What happens when labelling changes?



Future work

- Relationships between clique algorithm's runtime and properties of the association graph
- How the association graph changes after making a decision
- Can $k\downarrow$ and clique work together?