A Graph Database Repository and Performance Evaluation Metrics for Graph Edit Distance

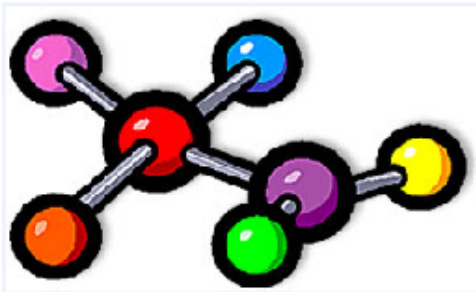
Zeina Abu-Aisheh, Romain Raveaux and Jean-Yves Ramel

Université de Tours - France
Equipe RFAI du Laboratoire Informatique (LI)

May 14, 2015

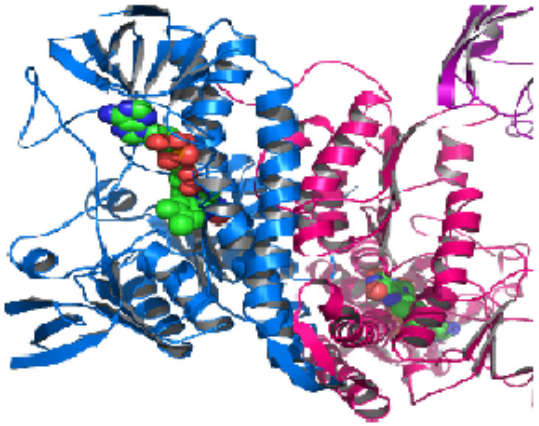
"Real-world graphs are everywhere" but how to compare them?

Social Networks



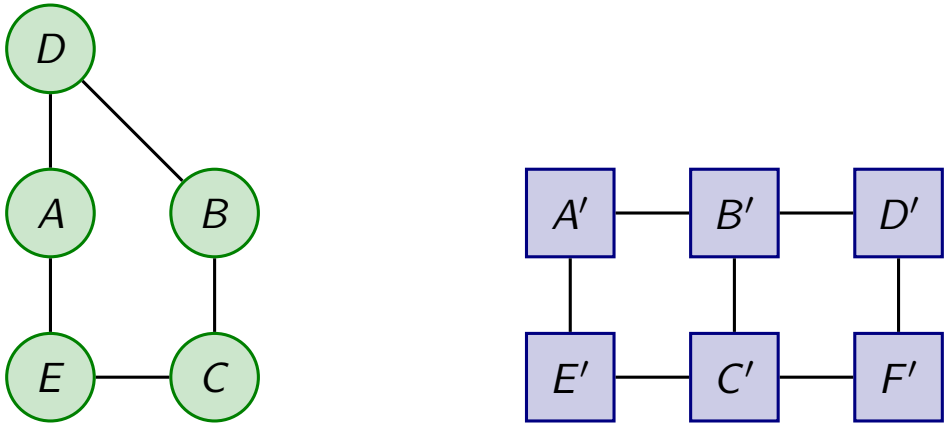
Chemical Bounds

Characters

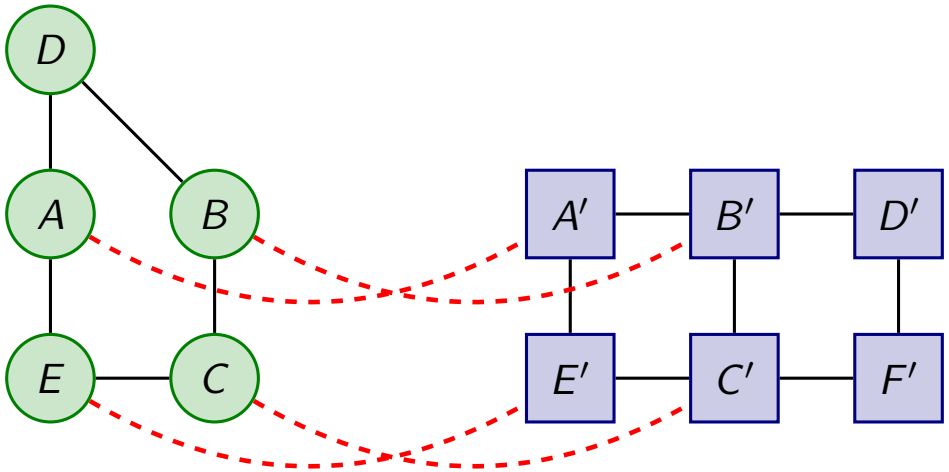


Protein Structures

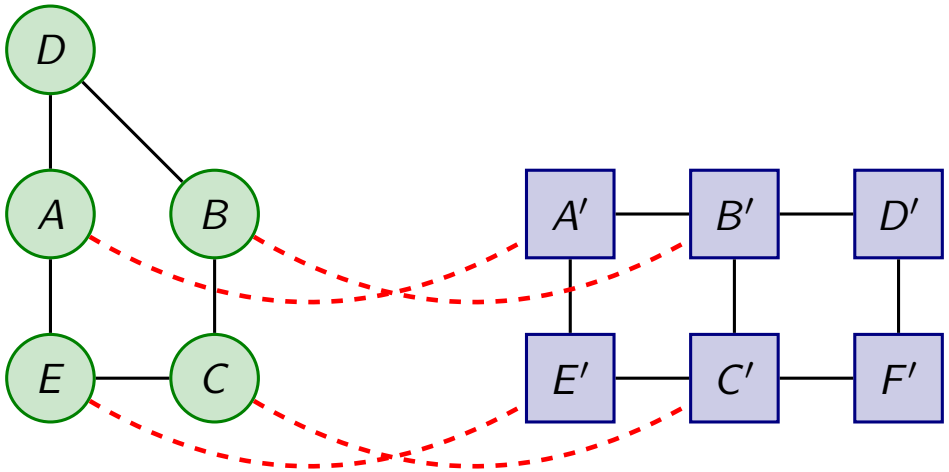
Inexact Graph Matching

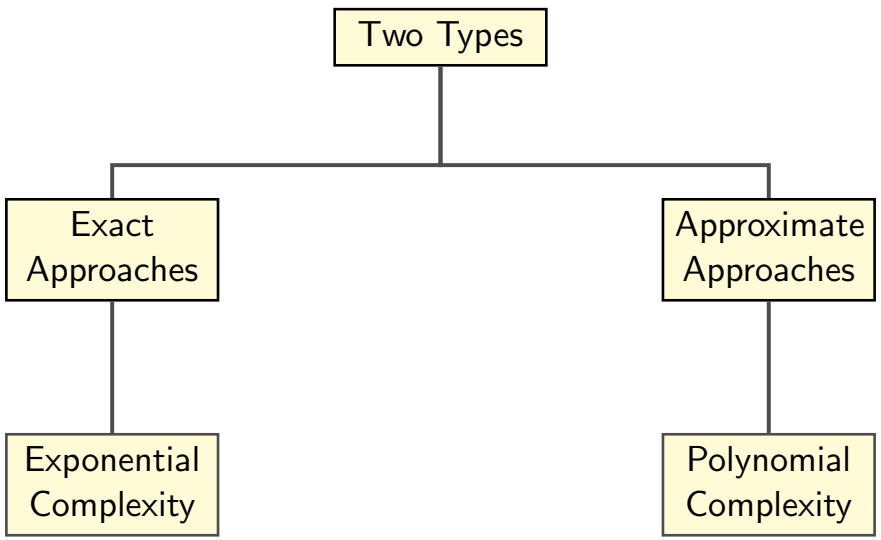


Inexact Graph Matching



Inexact Graph Matching





Raised Questions

Raised Questions

- **Q1: How to evaluate error tolerant graph comparison methods ?**

Raised Questions

- **Q1: How to evaluate error tolerant graph comparison methods ?**
  - Speed.

Raised Questions

- **Q1: How to evaluate error tolerant graph comparison methods ?**
  - Speed.
  - Precision.

Raised Questions

- **Q1: How to evaluate error tolerant graph comparison methods ?**
  - Speed.
  - Precision.
  - Memory consumption.

Raised Questions

- **Q1: How to evaluate error tolerant graph comparison methods ?**
  - Speed.
  - Precision.
  - Memory consumption.
  - Parameter's sensitivity.

Raised Questions

Raised Questions

- **Q2: How to choose graph matching databases?**

Raised Questions

- **Q2: How to choose graph matching databases?**
  - Graphs size.

Raised Questions

- **Q2: How to choose graph matching databases?**
  - Graphs size.
  - Graphs types:
    - ⋆ Different typologies (dense, sparse, etc.).
    - ⋆ Different kinds of attributes (nominal, numeric, etc.).

Raised Questions

- **Q2: How to choose graph matching databases?**
  - Graphs size.
  - Graphs types:
    - ⋆ Different typologies (dense, sparse, etc.).
    - ⋆ Different kinds of attributes (nominal, numeric, etc.).
  - Applications types:

Raised Questions

- **Q2: How to choose graph matching databases?**
  - Graphs size.
  - Graphs types:
    - ⋆ Different typologies (dense, sparse, etc.).
    - ⋆ Different kinds of attributes (nominal, numeric, etc.).
  - Applications types:
    - ⋆ Not only databases dedicated to classification (e.g., IAM [Riesen2008]).

Raised Questions

- **Q2: How to choose graph matching databases?**
  - Graphs size.
  - Graphs types:
    - ⋆ Different typologies (dense, sparse, etc.).
    - ⋆ Different kinds of attributes (nominal, numeric, etc.).
  - Applications types:
    - ⋆ Not only databases dedicated to classification (e.g., IAM [Riesen2008]).
    - ⋆ Other databases dedicated to other applications (e.g., matching quality).

Outline

Plan

Graph Edit Distance

Graph Edit Distance

Let $g_m = (V_m, E_m, \mu_m, \zeta_m)$ be the model graph and $g_t = (V_t, E_t, \mu_t, \zeta_t)$ be the target graph. The graph edit distance between G_1 and G_2 is defined by:

$$d_{plain}(g_m, g_t) = \min_{e_1, \cdots, e_k \in \gamma(g_1, g_2)} \sum_{i=1}^{k} c(e_i)$$

Figure: Graph Edit Distance operations

Related Works

Ref	Problem Type	Graph Type	Database Type	Measure Type	Purpose
Santo2003	Exact GM	Non-attributed	Synthetic	Accuracy and scalability	Matching
CMU DataBase	Error-tolerant GM	Attributed	Real-world	Memory consumption, accuracy and matching quality	Matching
Riesen2008	Error-tolerant GM	Attributed	Real-world	Accuracy and running time	Classification
Conte2007, Foggia2001	Exact GM	Attributed	Synthetic	Accuracy and scalability	Matching
Carletti2013	Exact GM	(Non)attributed	Real-world	Scalability	Matching

Table: Synthesis of graph databases.

CMU Database: `http://vasc.ri.cmu.edu/idb/html/motion`

Conclusions of the state of the art

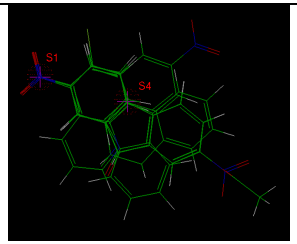
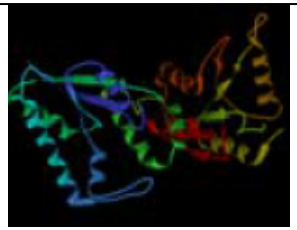
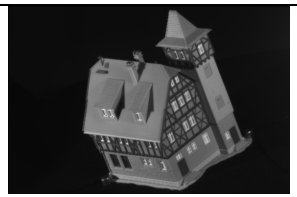
- Lack of performance comparison measures for error-tolerant GM methods.

Conclusions of the state of the art

- Lack of performance comparison measures for error-tolerant GM methods.
- None of the repositories was dedicated to assessing the performance of GED.

Plan

Databases

Database	Decomposition	Overview	Purpose
GREC	MIX, 5, 10, 15 and 20 vertices		Classification
MUTA	MIX, 10, 20, ... , 70 vertices		Classification
Protein	MIX, 20, 30 and 40 vertices		Classification
CMU	30 vertices		Matching

Databases

Database	# subsets	Max Graphs Size	# graphs per subset	# comparisons per subset	# optimal solutions
GREC	5 subsets	20	10 graphs	100	441
MUTA	8 subsets	70	10 graphs	100	189
Protein	5 subsets	40	10 graphs	100	47
CMU	1 subset	30	111 graphs	660	128

Table: Overview about the subsets included in the repository.

Repository link: `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.

Added Low Level Information

- **6 methods:**

Added Low Level Information

- **6 methods:**
  - Exact methods:

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - ⋆ A* [Riesen2007].

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - ⋆ A* [Riesen2007].
    - ⋆ DepthFirst-GED [Abu-Aisheh2015]: recent work published in ICPRAM2015.

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - ⋆ A* [Riesen2007].
    - ⋆ DepthFirst-GED [Abu-Aisheh2015]: recent work published in ICPRAM2015.
    - ⋆ Distributed DepthFirst-GED [AbuAisheh2015-under review].

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - A* [Riesen2007].
    - DepthFirst-GED [Abu-Aisheh2015]: recent work published in ICPRAM2015.
    - Distributed DepthFirst-GED [AbuAisheh2015-under review].
  - Approximate methods:

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - ⋆ A* [Riesen2007].
    - ⋆ DepthFirst-GED [Abu-Aisheh2015]: recent work published in ICPRAM2015.
    - ⋆ Distributed DepthFirst-GED [AbuAisheh2015-under review].
  - Approximate methods:
    - ⋆ BeamSearch [Neuhaus2006].

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - A* [Riesen2007].
    - DepthFirst-GED [Abu-Aisheh2015]: recent work published in ICPRAM2015.
    - Distributed DepthFirst-GED [AbuAisheh2015-under review].
  - Approximate methods:
    - BeamSearch [Neuhaus2006].
    - Munkres [Riesen2009].

Added Low Level Information

- **6 methods:**
  - Exact methods:
    - A* [Riesen2007].
    - DepthFirst-GED [Abu-Aisheh2015]: recent work published in ICPRAM2015.
    - Distributed DepthFirst-GED [AbuAisheh2015-under review].
  - Approximate methods:
    - BeamSearch [Neuhaus2006].
    - Munkres [Riesen2009].
    - Hausdorff [Fischer2013].

Added Low Level Information

- For each graph matching pair:

Added Low Level Information

- For each graph matching pair:
  - Distance between each pair of graphs.

Added Low Level Information

- For each graph matching pair:
  - Distance between each pair of graphs.
  - Vertex-Vertex matching.

Added Low Level Information

- For each graph matching pair:
  - Distance between each pair of graphs.
  - Vertex-Vertex matching.

Added Low Level Information

- For each graph matching pair:
  - Distance between each pair of graphs.
  - Vertex-Vertex matching.

G1Name	G2Name	Method	Distance	Optimal	Matching
image3_23	image3_25	*BS*−100	135.178	true	*Vertex*:0 → 0=37.476/ *Vertex*:1 → 1=6.519/ *Vertex*:2 → 2=32.070/ *Vertex*:4 → 4=34.409/ *Vertex*:3 → 3=24.703/ *Edge*:2 ↔ 3 → 2 ↔ 3 =0.0/ *Edge*:0 ↔ 4 → 0 ↔ 4=0.0

Table: Low level information (taken from the file GREC5.csv)

Added Low Level Information

- For each graph matching pair:
  - Distance between each pair of graphs.
  - Vertex-Vertex matching.

G1Name	G2Name	Method	Distance	Optimal	Matching
image3_23	image3_25	*BS*−100	135.178	true	*Vertex*:0 → 0=37.476/ *Vertex*:1 → 1=6.519/ *Vertex*:2 → 2=32.070/ *Vertex*:4 → 4=34.409/ *Vertex*:3 → 3=24.703/ *Edge*:2 ↔ 3 → 2 ↔ 3 =0.0/ *Edge*:0 ↔ 4 → 0 ↔ 4=0.0

Table: Low level information (taken from the file GREC5.csv)

- **Detailed answers:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.

Plan

Performance Evaluation Metrics

- All metrics can be tested under:

Performance Evaluation Metrics

- All metrics can be tested under:
  - Time Constraint (C_T)

Performance Evaluation Metrics

- All metrics can be tested under:
  - Time Constraint (C_T)
  - Memory Constraint (C_M)

Performance Evaluation Metrics

Performance Evaluation Metrics

- Given the reference answer R:

Performance Evaluation Metrics

- Given the reference answer R:
  - Deviation.

Performance Evaluation Metrics

- Given the reference answer R:
  - Deviation.
  - Matching Dissimilarity.

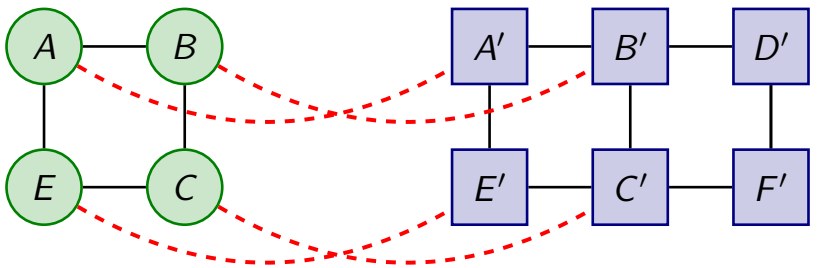
Performance Evaluation Metrics

- Given the reference answer R:
  - Deviation.
  - Matching Dissimilarity.

Performance Evaluation Metrics

- Given the reference answer R:
  - Deviation.
  - Matching Dissimilarity.

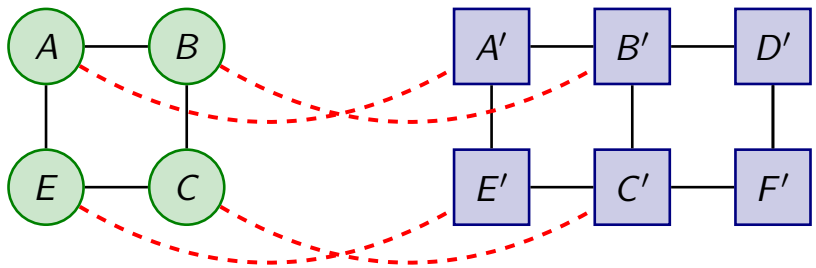
- Reference Distance (RD) = 5
- Reference Matching (RM):



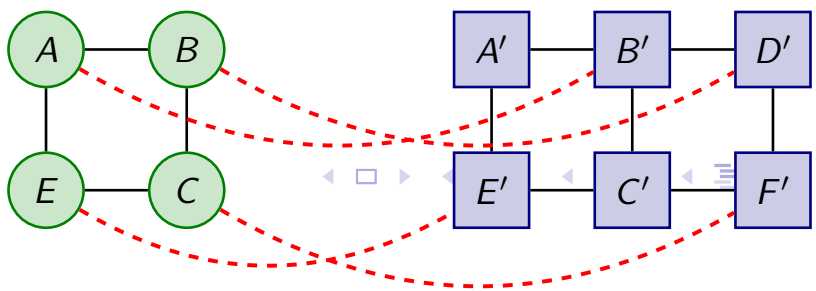
Performance Evaluation Metrics

- Given the reference answer R:
  - Deviation.
  - Matching Dissimilarity.

- Reference Distance (RD) = 5
- Reference Matching (RM):



- Distance found by method (m) = 6.
  - Deviation = 20%.
- Matching found by method (m):
  - Matching Dissimilarity = 75%.



Performance Evaluation Metrics

- For a set S of GED methods:

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.
    - $\star$ i.e. *Incomplete* editpath.

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.
    - $\star$ i.e. *Incomplete* editpath.
  - Number of Time-Out and Out-Of-Memory cases.

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.
    - $\star$ i.e. *Incomplete* editpath.
  - Number of Time-Out and Out-Of-Memory cases.
  - Mean number of explored nodes.

Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.
    - $\star$ i.e. *Incomplete* editpath.
  - Number of Time-Out and Out-Of-Memory cases.
  - Mean number of explored nodes.
  - Mean running time in milliseconds.

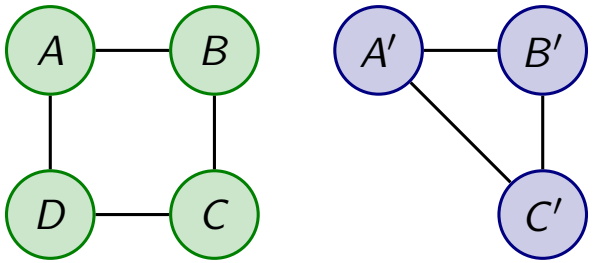
Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.
    - $\star$ i.e. *Incomplete* editpath.
  - Number of Time-Out and Out-Of-Memory cases.
  - Mean number of explored nodes.
  - Mean running time in milliseconds.
  - Running time-Deviation plot:

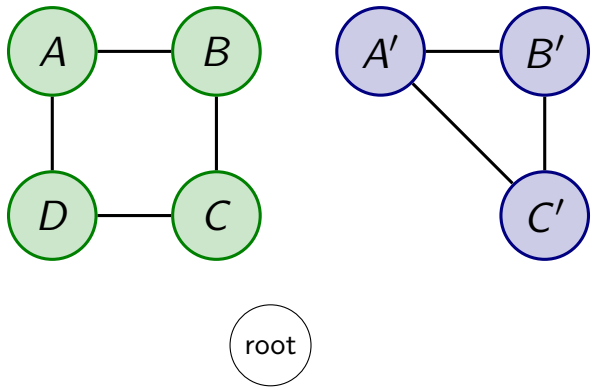
Performance Evaluation Metrics

- For a set S of GED methods:
  - Number of best found solutions:
    - $\star$ # times a method $m \in S$ found the *best* solution.
  - Number of optimal solutions:
    - $\star$ # times a method $m \in S$ found the *optimal* solution.
  - Number of unfeasible solutions.
    - $\star$ i.e. *Incomplete* editpath.
  - Number of Time-Out and Out-Of-Memory cases.
  - Mean number of explored nodes.
  - Mean running time in milliseconds.
  - Running time-Deviation plot:
    - $\star$ Projection on a two-dimensional space ($\mathbb{R}^2$).

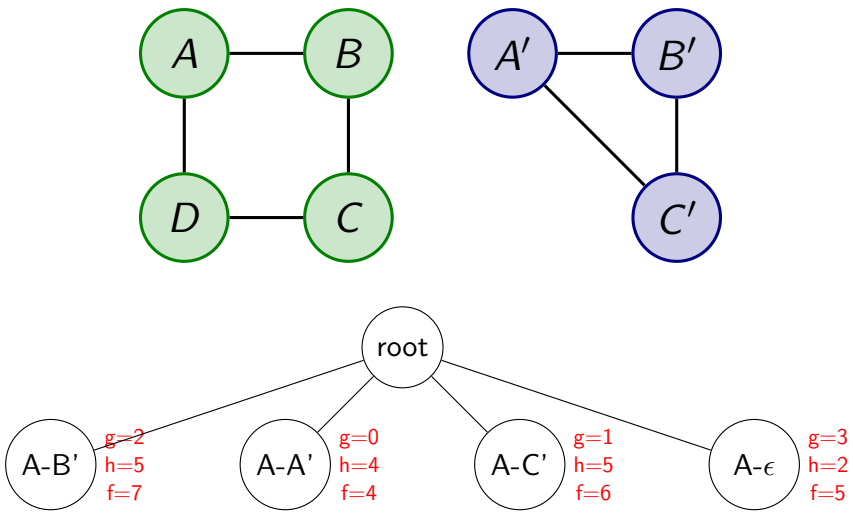
Performance Evaluation Metrics - Example



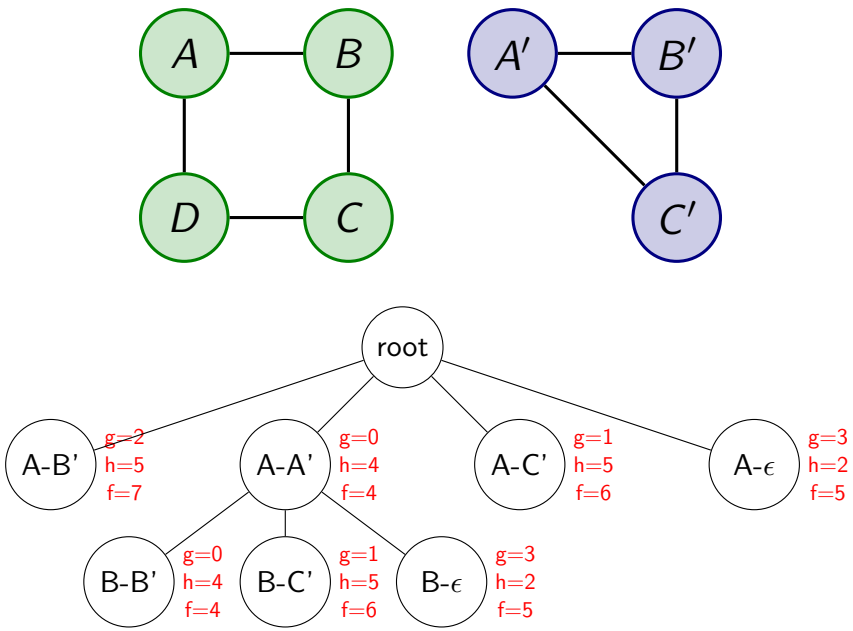
Performance Evaluation Metrics - Example



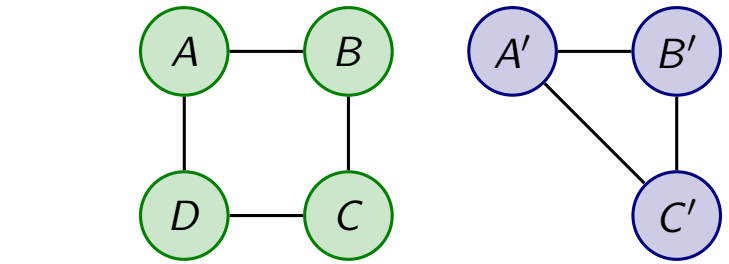
Performance Evaluation Metrics - Example

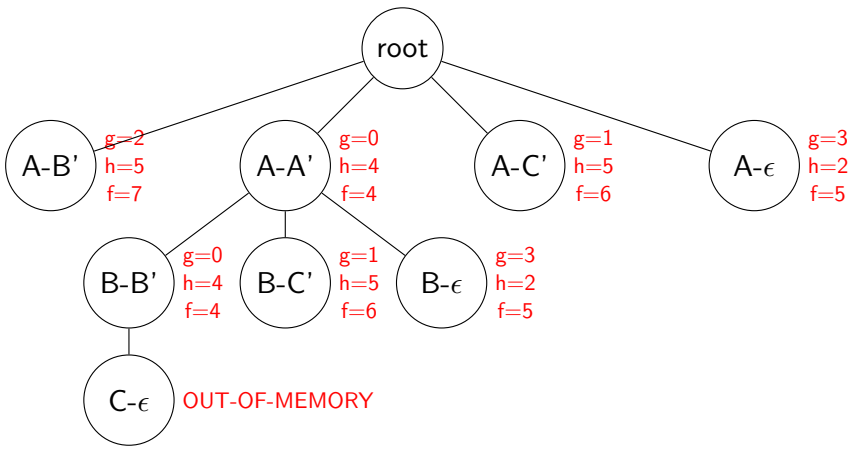


Performance Evaluation Metrics - Example

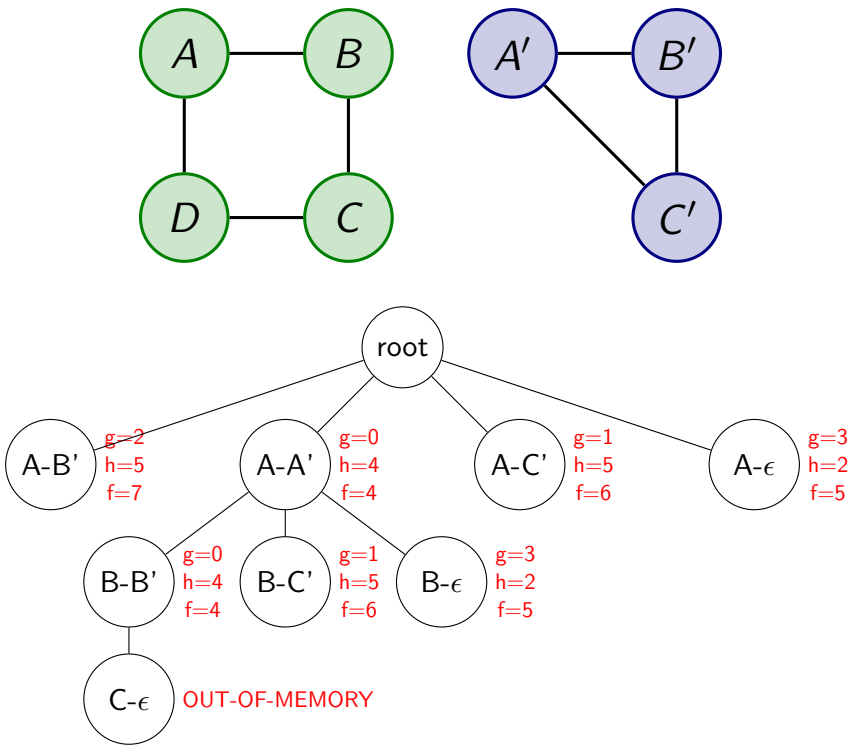


Performance Evaluation Metrics - Example





Performance Evaluation Metrics - Example



- Out-of-memory.
- Unfeasible solution.
- # of explored nodes = 3.

Plan

Use Case

- **Compared Methods**:

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].
  - Approximate GED: BeamSearch (BS) [Neuhaus2006].

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].
  - Approximate GED: BeamSearch (BS) [Neuhaus2006].
    - $\star$ $x = 1$ and 100

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].
  - Approximate GED: BeamSearch (BS) [Neuhaus2006].
    - $x = 1$ and 100
- **Database:** Mutagenicity.

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].
  - Approximate GED: BeamSearch (BS) [Neuhaus2006].
    - $x = 1$ and 100
- **Database:** Mutagenicity.
- **Constraints:**

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].
  - Approximate GED: BeamSearch (BS) [Neuhaus2006].
    - $x = 1$ and 100
- **Database:** Mutagenicity.
- **Constraints:**
  - $C_T = 300$ seconds.

Use Case

- **Compared Methods**:
  - Exact GED: A* [Riesen2007].
  - Approximate GED: BeamSearch (BS) [Neuhaus2006].
    - $x = 1$ and 100
- **Database:** Mutagenicity.
- **Constraints:**
  - $C_T = 300$ seconds.
  - $C_M = 1024$ MB.

Results - Comparing with Reference Answers

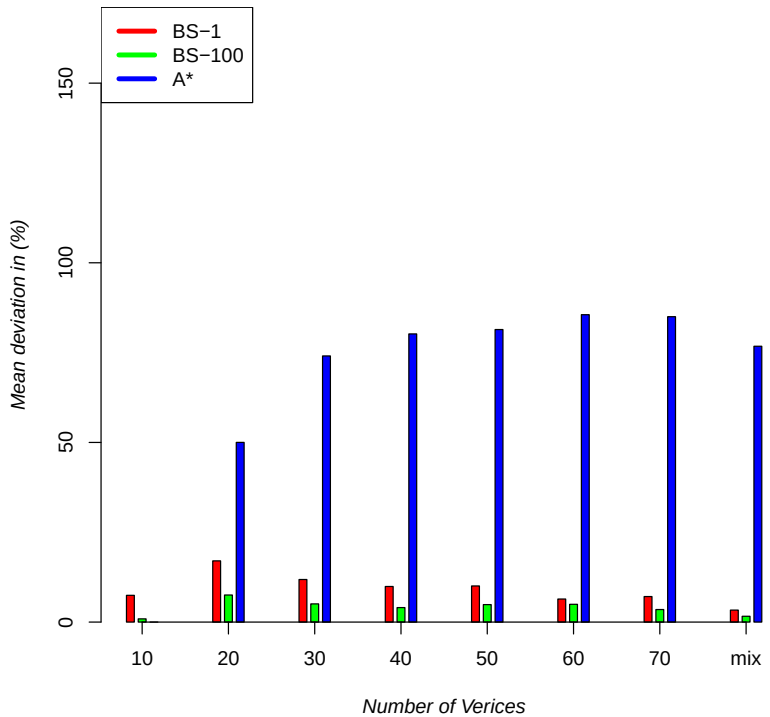


Figure: Deviation

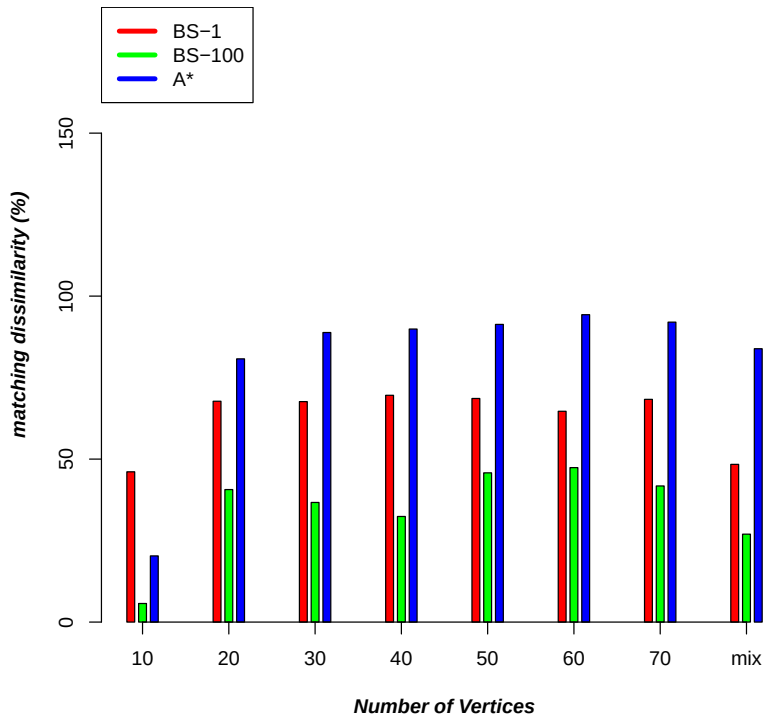


Figure: Matching Dissimilarity

The closer to 0, the better

Results - General Information

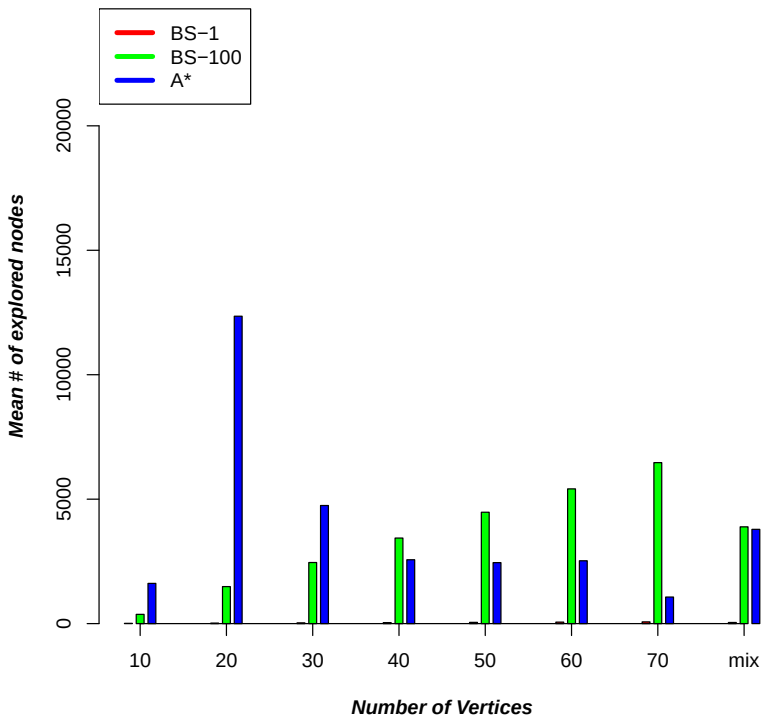


Figure: No of Explored Nodes

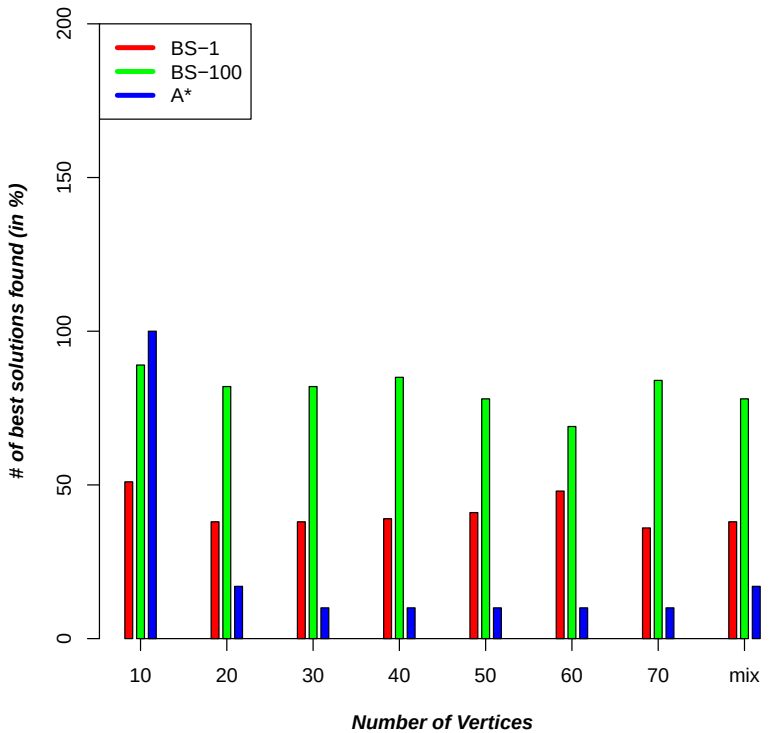


Figure: No of Best Found Solutions

Results - General Information

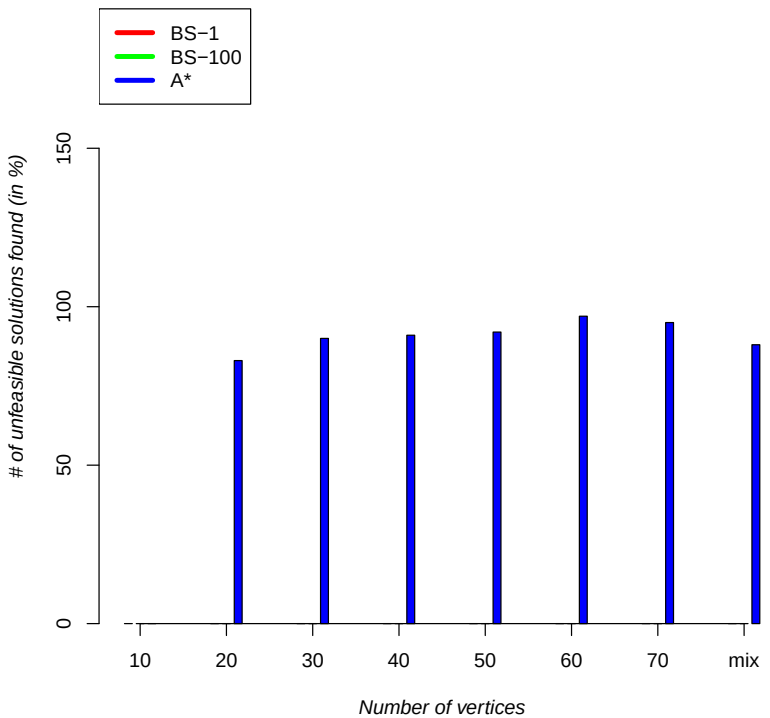


Figure: No of Unfeasible Solutions

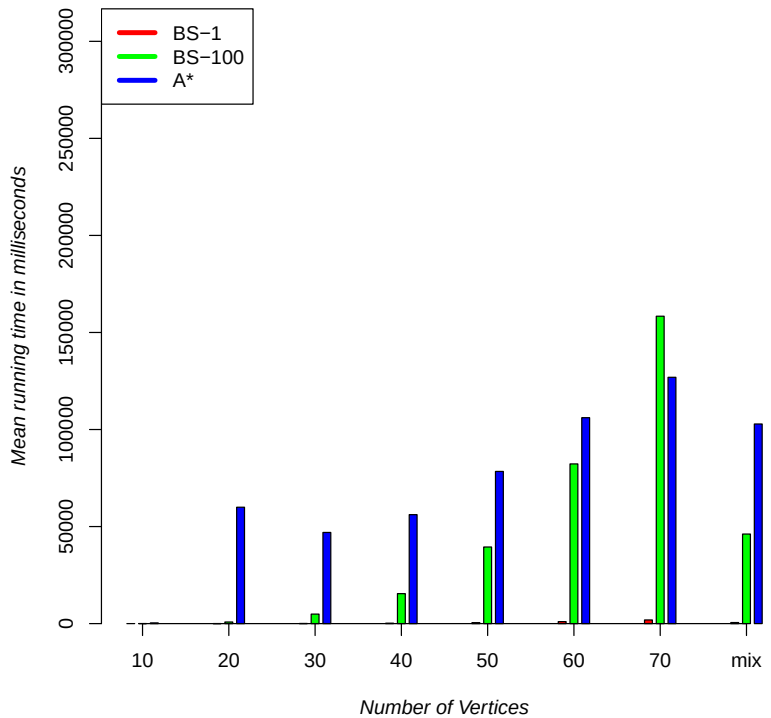


Figure: Running Time

Results - General Information

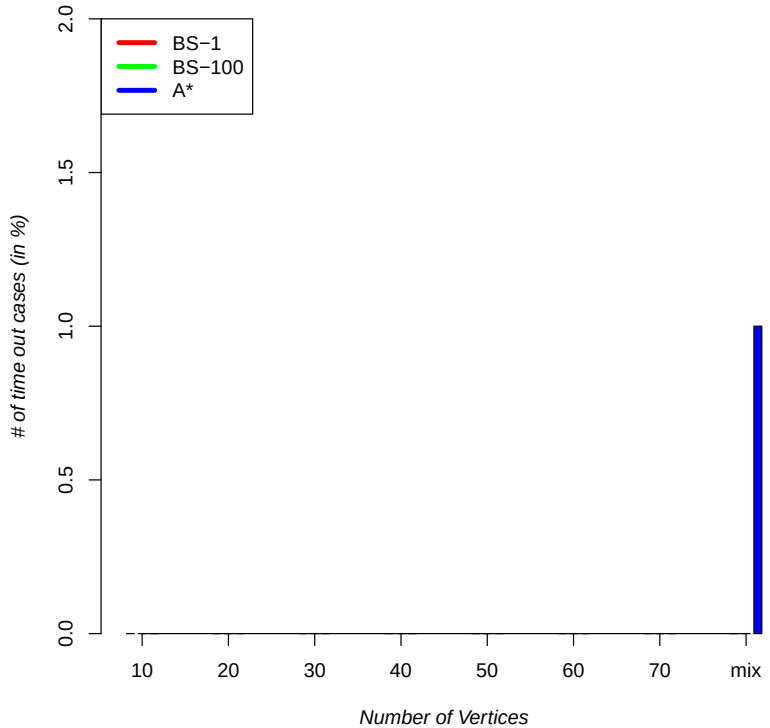


Figure: No of Time-Outs

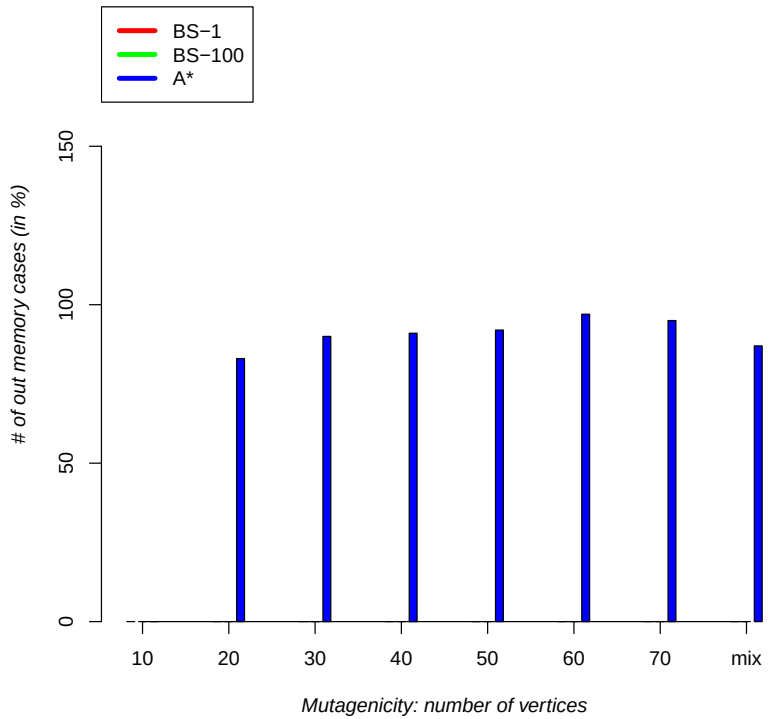
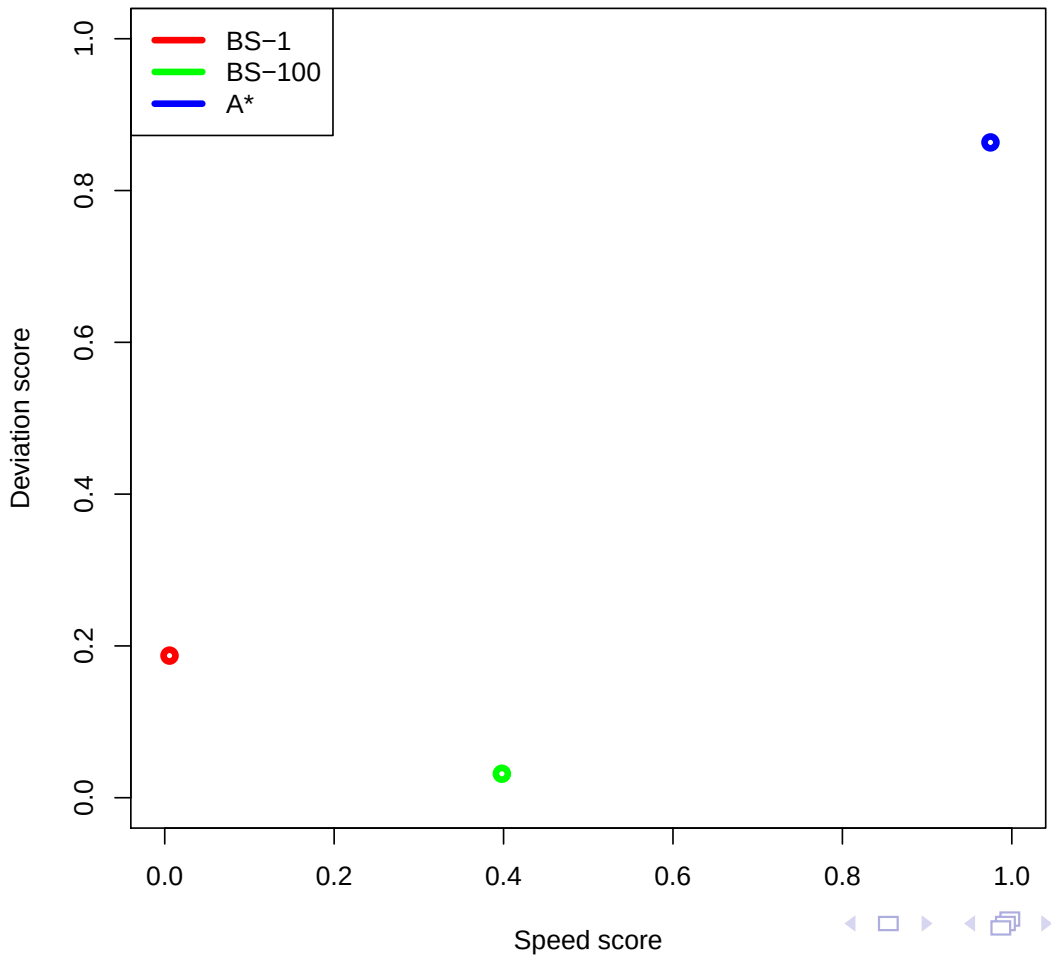


Figure: No of Memory-Outs

Results - Plot



Plan

First Contribution

- **Public repository:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.

First Contribution

- **Public repository:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.
- **Additional low level annotation:**

First Contribution

- **Public repository:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.
- **Additional low level annotation:**
  - Best distance and vertex-vertex matchings.

First Contribution

- **Public repository:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.
- **Additional low level annotation:**
  - Best distance and vertex-vertex matchings.
  - 6 methods (3 exact and 3 approximate).

First Contribution

- **Public repository:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.
- **Additional low level annotation:**
  - Best distance and vertex-vertex matchings.
  - 6 methods (3 exact and 3 approximate).
  - Four databases (GREC, Mutagenicity, Protein and CMU).

First Contribution

- **Public repository:** `http://www.rfai.li.univ-tours.fr/PagesPerso/zabuaisheh/GED-benchmark.html`.
- **Additional low level annotation:**
  - Best distance and vertex-vertex matchings.
  - 6 methods (3 exact and 3 approximate).
  - Four databases (GREC, Mutagenicity, Protein and CMU).
  - **Scalability:** database decomposition.

Second Contribution

- **Performance evaluation metrics:**

Second Contribution

- **Performance evaluation metrics:**
  - Under time and memory constraints.

Second Contribution

- **Performance evaluation metrics:**
  - Under time and memory constraints.
  - **Metrics:**

Second Contribution

- **Performance evaluation metrics:**
  - Under time and memory constraints.
  - **Metrics:**
    - ⋆ Deviation.

Second Contribution

- **Performance evaluation metrics:**
  - Under time and memory constraints.
  - **Metrics:**
    - ⋆ Deviation.
    - ⋆ Speed.

Second Contribution

- **Performance evaluation metrics:**
  - Under time and memory constraints.
  - **Metrics:**
    - ⋆ Deviation.
    - ⋆ Speed.
    - ⋆ Matching Dissimilarity.

Second Contribution

- **Performance evaluation metrics:**
  - Under time and memory constraints.
  - **Metrics:**
    - ⋆ Deviation.
    - ⋆ Speed.
    - ⋆ Matching Dissimilarity.
    - ⋆ Deviation-Running time plot.

Second Contribution

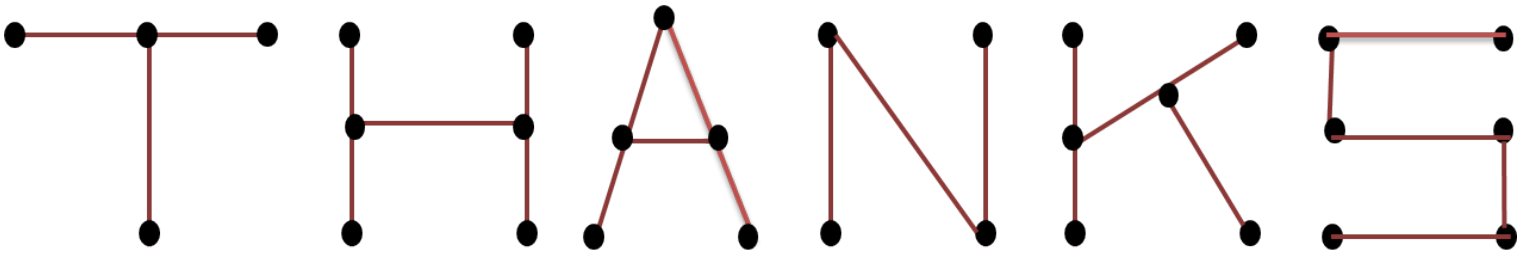
- **Performance evaluation metrics:**
  - Under time and memory constraints.
  - **Metrics:**
    - ⋆ Deviation.
    - ⋆ Speed.
    - ⋆ Matching Dissimilarity.
    - ⋆ Deviation-Running time plot.
    - ⋆ Number of optimal solutions.

Perspectives

- Integrating better answers in the database (if found).

Perspectives

- Integrating better answers in the database (if found).
- Expanding this repository by integrating other publicly available databases.


THANKS