Comparing heuristics for graph edit distance computation

D. B. Blumenthal[1, 2, ✉], N. Boria[3], J. Gamper[1], S. Bougleux[3], L. Brun[3]

[1]Free University of Bozen-Bolzano, Faculty of Computer Science, Bolzano, Italy
[2]Technical University of Munich, Chair of Experimental Bioinformatics, Freising, Germany
[3]Normandie Université, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France
[✉]david.blumenthal@wzw.tum.de, david.blumenthal@inf.unibz.it

https://github.com/dbblumenthal/gedlib

Graph Edit Distance (Definition)

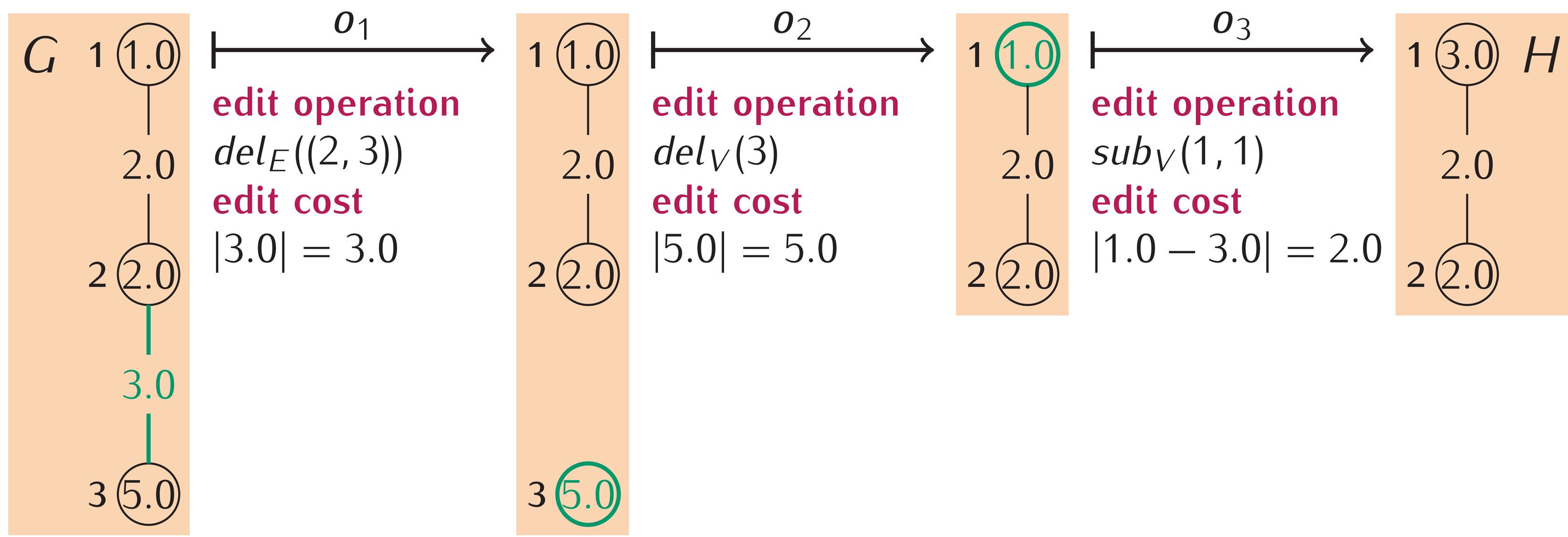
- $\mathrm{GED}(G,H) := \min\{c(P) \mid P \text{ edit path between } G \text{ and } H\}$
- labeled graphs: $G = (V^G, E^G, \ell_V^G, \ell_E^G)$ and $H = (V^H, E^H, \ell_V^H, \ell_E^H)$
- labeling functions: $\ell_V^G : V^G \to \Sigma_V$ and $\ell_E^G : E^G \to \Sigma_E$
- node and edge label spaces: Σ_V and Σ_E
- edit path: edit operation sequence $P = (o_1, \ldots, o_r)$ transforms G into H
- edit operations and edit costs ($i \in V^G$, $k \in V^H$, $(i,j) \in V^G$, $(k,l) \in E^H$):

node edit op.	edit cost	edge edit op.	edit cost
$sub_V(i,k)$	$c_V(\ell_V^G(i), \ell_V^H(k))$	$sub_E((i,j),(k,l))$	$c_E(\ell_E^G(i,j), \ell_E^H(k,l))$
$del_V(i)$	$c_V(\ell_V^G(i), \epsilon)$	$del_E((i,j))$	$c_E(\ell_E^G(i,j), \epsilon)$
$ins_V(k)$	$c_V(\epsilon, \ell_V^H(k))$	$ins_E((k,l))$	$c_E(\epsilon, \ell_E^H(k,l))$

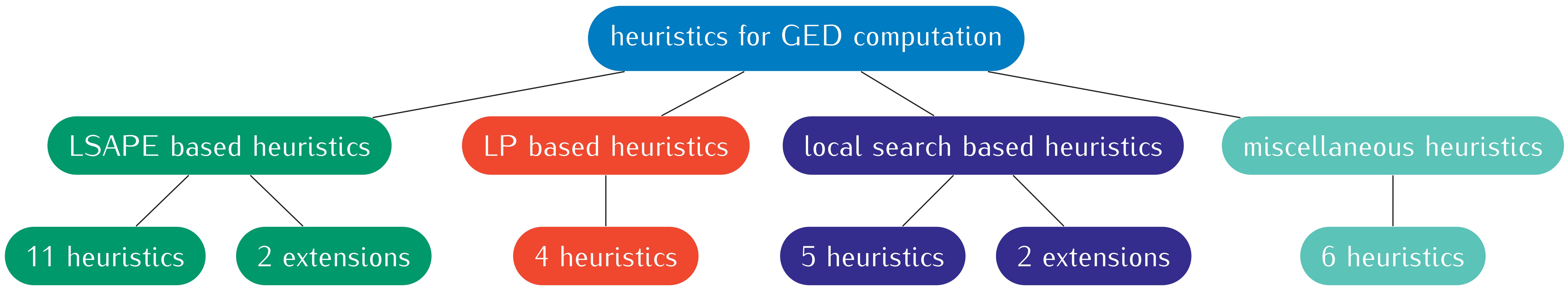
- edit path cost: $c(P) = \sum_{i=1}^{r} c(o_i)$

Graph Edit Distance (Example)

- node and edge label spaces: $\Sigma_V = \Sigma_E = \mathbb{R}_{\geq 0}$
- sub. costs: $c_V(\alpha,\beta) = c_E(\alpha,\beta) = |\alpha - \beta|$
- del. and ins. costs: $c_V(\alpha,\epsilon) = c_V(\epsilon,\alpha) = c_E(\alpha,\epsilon) = c_E(\epsilon,\alpha) = |\alpha|$
- edit path cost: $c(P) = c(o_1) + c(o_2) + c(o_3) = 3.0 + 5.0 + 2.0 = 10.0$



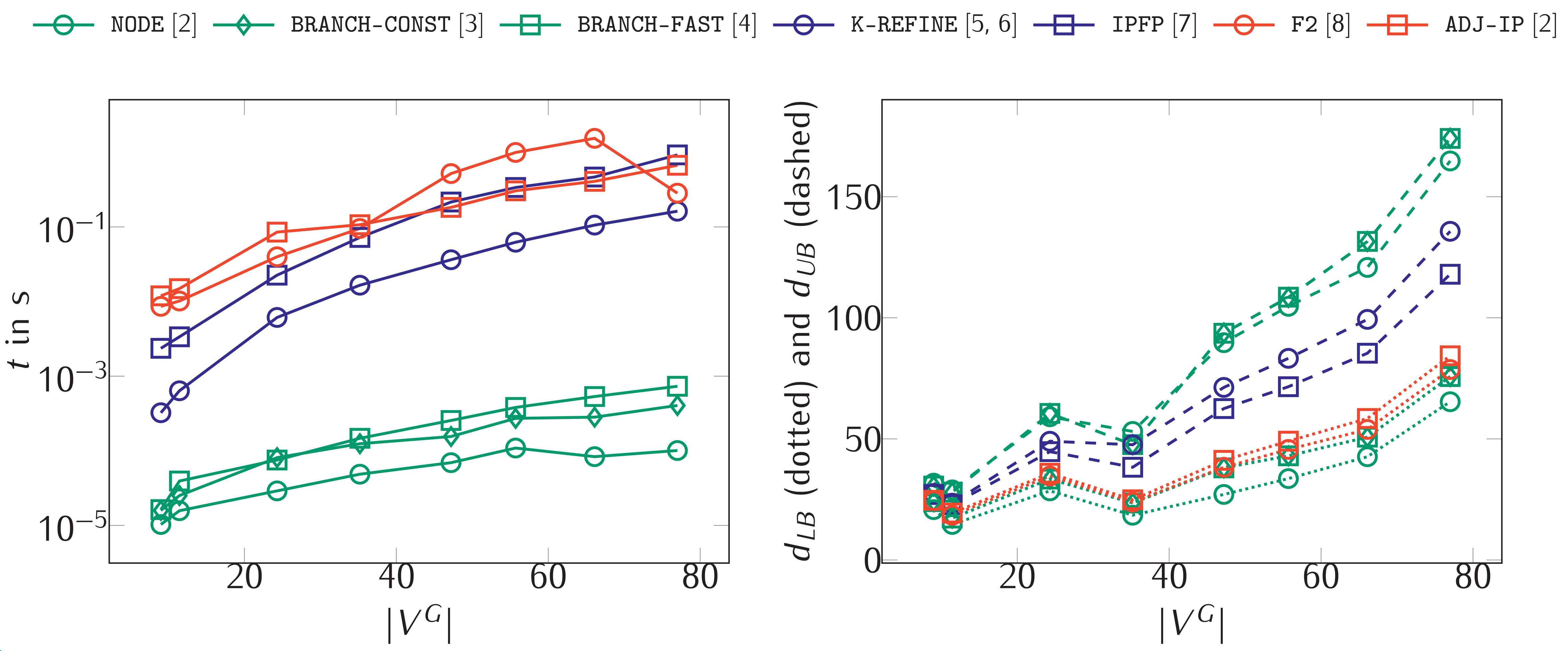
Suggested Taxonomy and Compared Heuristics



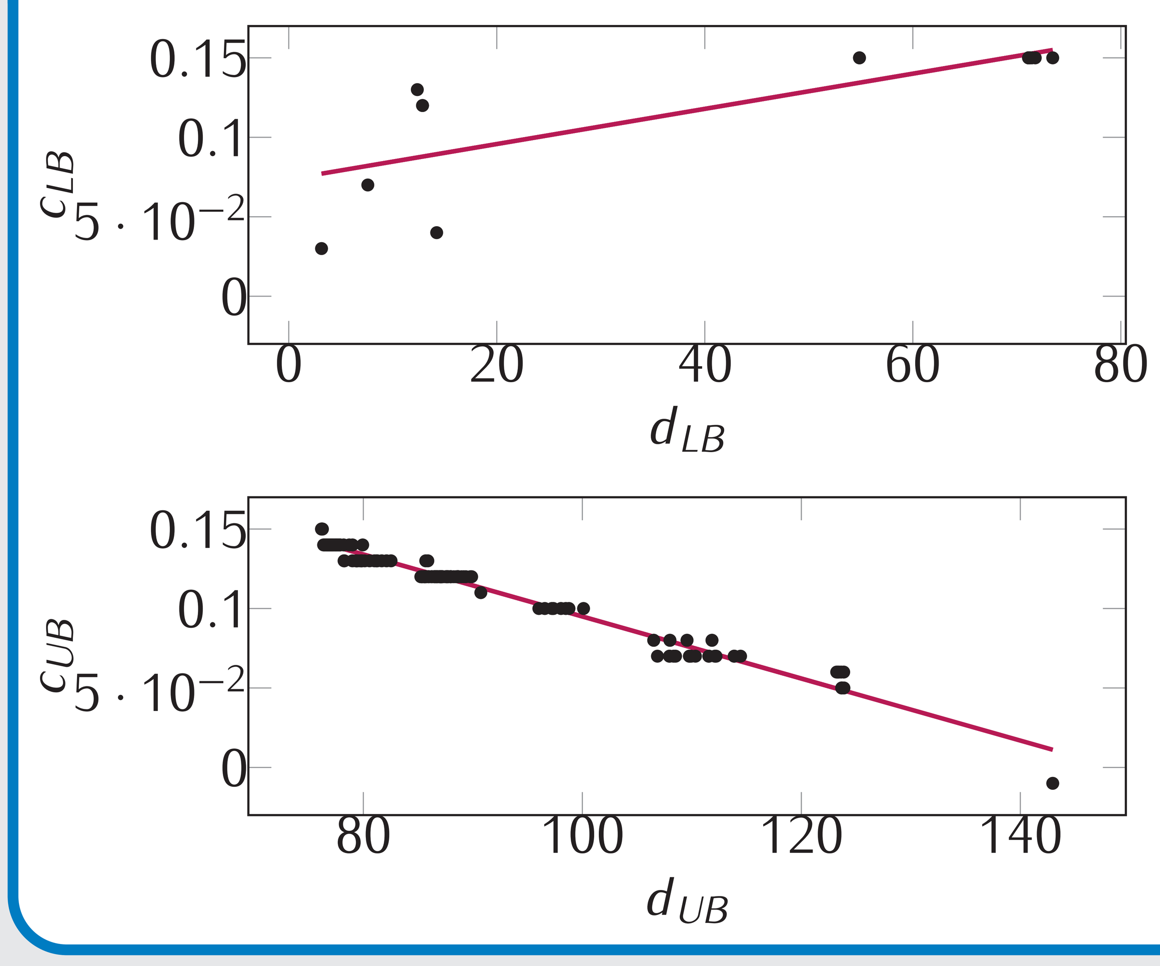
Experimental Setup

- datasets: 6 widely used benchmark datasets from IAM Graph Database Repository [1]
- implementation: to ensure comparability, all heuristics were re-implemented in C++
- metrics: runtime t, lower and upper bounds d_{LB} and d_{UB}, and classification coefficients c_{LB} and c_{UB}

Runtimes and Bounds of Globally Best Heuristics



Bounds vs. Classif. Coeff.



Main Insights

- tightness of lower bound critical: use LP based heuristic F2 [8] or ADJ-IP [2]
- tightness of upper bound critical: use local search based heuristic IPFP [7] or K-REFINE [5, 6]
- runtime critical: use LSAPE based heuristic NODE [2], BRANCH-CONST [3], or BRANCH-FAST [4]
- stability w. r. t. graph size: ordering of heuristics w. r. t. tightness is independent of graph size
- confirmed hypothesis: tight bounds exhibit better classification coefficients than loose bounds
- gap between best lower and upper bounds: only grows moderately with increasing graph size

Presented at VLDB 2019, Los Angeles, USA, August 26–30, 2019

References

[1] K. Riesen and H. Bunke, “IAM graph database repository for graph based pattern recognition and machine learning,” in *S+SSPR*, 2008, pp. 287–297.

[2] D. Justice and A. Hero, “A binary linear programming formulation of the graph edit distance,” *TPAMI*, vol. 28, no. 8, pp. 1200–1214, 2006.

[3] W. Zheng, L. Zou, *et al.*, “Efficient graph similarity search over large graph databases,” *TKDE*, vol. 27, no. 4, pp. 964–978, 2015.

[4] D. B. Blumenthal and J. Gamper, “Improved lower bounds for graph edit distance,” *TKDE*, vol. 30, no. 3, pp. 503–516, 2018.

[5] Z. Zeng, A. K. H. Tung, *et al.*, “Comparing stars: On approximating graph edit distance,” *PVLDB*, vol. 2, no. 1, pp. 25–36, 2009.

[6] N. Boria, D. B. Blumenthal, *et al.*, “Improved local search for graph edit distance,” 2019. arXiv: 1907.02929 [cs.DS].

[7] S. Bougleux, B. Gaüzère, *et al.*, “Graph edit distance as a quadratic program,” in *ICPR*, 2016, pp. 1701–1706.

[8] J. Lerouge, Z. Abu-Aisheh, *et al.*, “New binary linear programming formulation to compute the graph edit distance,” *PR*, vol. 72, pp. 254–265, 2017.