

Table 1: Performance metrics for different algorithms and models. Results with † are run with our code; with \* copied from related work. Models with a diamond (◊) iterate for the correct amount of steps during train time (may differ between datapoints). Filled diamond (◆) means the ground truth number of steps is also given at test time. HCS stands for **h**ard**c**oded **s**teps to 64 during test time; colour is used if on average 64 steps are **more/less** than ground truth. LT stands for **l**earnt **t**ermination.

| Algorithm              | NAR <sup>†◆</sup> | NAR <sup>†◊</sup><br>(HCS) | NAR <sup>*◆</sup><br>(Triplet-MPNN) | NAR <sup>†◆</sup><br>(Triplet-MPNN) | NAR <sup>†◊</sup><br>(LT) | DEAR <sup>†</sup><br>(ours) |
|------------------------|-------------------|----------------------------|-------------------------------------|-------------------------------------|---------------------------|-----------------------------|
| <b>Bellman-F.</b>      | 97.06% ± 0.40     | <b>68.26% ± 13.58</b>      | 93.26% ± 0.04                       | 97.23% ± 0.15                       | 95.39% ± 1.42             | 96.78% ± 0.43               |
| <b>Floyd-W.</b>        | 52.53% ± 0.98     | 52.53% ± 0.98              | 40.80% ± 2.90                       | 61.86% ± 1.57                       | 49.30% ± 0.53             | 55.75% ± 2.20               |
| <b>DSP</b>             | 94.21% ± 1.77     | <b>93.56% ± 1.16</b>       | 97.62% ± 0.21                       | 93.32% ± 1.60                       | 88.30% ± 1.04             | 89.81% ± 0.14               |
| <b>MST Prim</b>        | 93.56% ± 0.77     | 93.56% ± 0.77              | 99.24% ± 0.21                       | 92.01% ± 1.50                       | 87.69% ± 1.17             | 88.67% ± 0.74               |
| <b>BFS</b>             | 99.85% ± 0.09     | <b>87.42% ± 11.91</b>      | 99.89% ± 0.03                       | 99.69% ± 0.29                       | 99.51% ± 0.06             | 98.73% ± 0.37               |
| <b>DFS</b>             | 16.89% ± 5.73     | <b>27.54% ± 0.52</b>       | 47.79% ± 4.19                       | 31.20% ± 4.02                       | 29.07% ± 2.32             | 40.62% ± 0.44               |
| <b>SCC</b>             | 40.70% ± 1.39     | <b>41.79% ± 0.56</b>       | 57.63% ± 0.68                       | 46.84% ± 1.70                       | 39.33% ± 1.52             | 43.63% ± 1.19               |
| <b>Search (Binary)</b> | 94.67% ± 2.31     | <b>27.00% ± 24.25</b>      | 93.21% ± 1.10                       | 93.33% ± 2.31                       | 84.33% ± 8.33             | 59.00% ± 12.3               |
| <b>Minimum</b>         | 97.67% ± 5.73     | 97.67% ± 5.73              | 99.24% ± 0.21                       | 96.67% ± 2.31                       | 94.00% ± 2.00             | 97.22% ± 3.82               |
| <b>Sort (Ins.)</b>     | 27.07% ± 10.3     | 27.07% ± 10.3              | 77.29% ± 7.42                       | 63.67% ± 39.97                      | 33.8% ± 12.04             | 86.93% ± 3.87               |
| <b>Overall</b>         | 71.42%            | 61.64%                     | 80.60%                              | 77.58%                              | 70.07%                    | 75.42%                      |

Table 2: Fixing anomalies with CLRS-30’s binary search further increases our overall score making our approach competitive to Triplet-MPNN. In the tables below, we will use the fixed version of search (which we reran for all models). Notation taken from Table 1.

| Algorithm          | NAR <sup>†◆</sup> | NAR <sup>†◊</sup><br>(HCS) | NAR <sup>†◆</sup><br>(Triplet-MPNN) | NAR <sup>†◊</sup><br>(LT) | DEAR <sup>†</sup><br>(ours) |
|--------------------|-------------------|----------------------------|-------------------------------------|---------------------------|-----------------------------|
| <b>Search</b>      | 95.67% ± 0.58     | <b>95.00% ± 1.73</b>       | 93.33% ± 0.58                       | 93.33% ± 3.05             | 85.67% ± 0.58               |
| <b>New Overall</b> | 71.52%            | 68.44%                     | 77.58%                              | 70.97%                    | 78.38%                      |

Table 3: Extreme OOD testing of NAR and DEAR. Search may skew the results at larger scales, hence we report the overall without it too.

| Algorithm                 | Size 128 (8×)     |                             | Size 256 (16×)    |                             | Size 512 (32×)    |                             |
|---------------------------|-------------------|-----------------------------|-------------------|-----------------------------|-------------------|-----------------------------|
|                           | NAR <sup>†◆</sup> | DEAR <sup>†</sup><br>(ours) | NAR <sup>†◆</sup> | DEAR <sup>†</sup><br>(ours) | NAR <sup>†◆</sup> | DEAR <sup>†</sup><br>(ours) |
| <b>Bellman-F.</b>         | 94.98% ± 0.52     | 95.65% ± 0.45               | 91.83% ± 1.50     | 93.09% ± 1.33               | 87.29% ± 1.89%    | 89.05% ± 2.11%              |
| <b>Floyd-W.</b>           | 28.16% ± 1.38     | 31.37% ± 2.59               | 39.28% ± 0.71     | 41.63% ± 1.61               | OOM               | OOM                         |
| <b>DSP</b>                | 88.52% ± 3.44     | 81.21% ± 4.29               | 78.65% ± 5.04     | 64.72% ± 8.73               | 54.38% ± 17.71%   | 43.98% ± 9.75%              |
| <b>MST Prim</b>           | 86.95% ± 1.11     | 80.59% ± 1.76               | 75.83% ± 2.69%    | 69.59% ± 3.13               | 66.18% ± 4.08%    | 59.83% ± 3.23%              |
| <b>BFS</b>                | 99.95% ± 0.00     | 98.87% ± 0.44               | 99.62% ± 0.12     | 97.35% ± 0.05               | 99.71% ± 0.21%    | 97.52% ± 0.87%              |
| <b>DFS</b>                | 8.14% ± 2.04      | 21.67% ± 1.64               | 3.65% ± 0.75      | 11.53% ± 1.49               | 1.24% ± 0.20%     | 4.47% ± 0.63%               |
| <b>SCC</b>                | 22.93% ± 1.48     | 28.56% ± 1.25               | OOM               | OOM                         | OOM               | OOM                         |
| <b>Search</b>             | 87.67% ± 5.51     | 71.67% ± 11.37              | 79.67% ± 10.02    | 42.67% ± 9.71%              | 72.00% ± 22.07%   | 20.00% ± 5.29%              |
| <b>Minimum</b>            | 94.00% ± 0.00     | 94.00% ± 3.46               | 92.33% ± 0.58     | 91.67% ± 3.06               | 87.00% ± 1.0%     | 86.67% ± 2.89%              |
| <b>Sort (Ins.)</b>        | 11.59% ± 4.85     | 48.90% ± 11.41              | 3.53% ± 1.11      | 16.37% ± 10.5               | 1.63% ± 0.71%     | 5.83% ± 4.80%               |
| <b>Overall</b>            | 62.29%            | 65.25%                      | 62.71%            | 58.74%                      | 58.68%            | 50.92%                      |
| <b>Overall w/o search</b> | 59.47%            | 64.54%                      | 60.59%            | 60.75%                      | 56.78%            | 55.34%                      |

Table 4: Mean runtime per sample in seconds (A100 80GB GPU). OOM indicates out-of-memory. Rows and columns correspond to Table 3. Vertical bar | is used to separate the symbols (up/down arrows) across the three sizes. Double symbol is used for substantial differences.

| Algorithm          | Size 128 (8×) |      |   | Size 256 (16×) |        | Size 512 (32×) |        |
|--------------------|---------------|------|---|----------------|--------|----------------|--------|
|                    | NAR           | DEAR |   | NAR            | DEAR   | NAR            | DEAR   |
| <b>Bellman-F.</b>  | ↑             | ↑    | ↑ | 0.0340         | 0.0575 | 0.0413         | 0.0912 |
| <b>Floyd-W.</b>    | ↓             | ↓    | — | 0.5856         | 0.1578 | 2.9694         | 0.3850 |
| <b>DSP</b>         | ↓             | ↓    | ↓ | 0.6756         | 0.1339 | 1.5654         | 0.1632 |
| <b>MST Prim</b>    | ↓             | ↓    | ↓ | 0.3883         | 0.0766 | 0.9941         | 0.1203 |
| <b>BFS</b>         | ↑             | ↑    | ↑ | 0.0223         | 0.0310 | 0.0283         | 0.0380 |
| <b>DFS</b>         | ↓             | ↓    | ↓ | 1.2965         | 0.1186 | 3.8432         | 0.1511 |
| <b>SCC</b>         | ↓             | —    | — | 2.1008         | 0.0908 | OOM            | OOM    |
| <b>Search</b>      | ↑             | ≈    | ≈ | 0.0365         | 0.0426 | 0.0725         | 0.0787 |
| <b>Minimum</b>     | ↓             | ↓    | ↓ | 0.3451         | 0.0413 | 1.3009         | 0.0798 |
| <b>Sort (Ins.)</b> | ↓             | ↓    | ↓ | 2.9813         | 0.0643 | 6.9552         | 0.1100 |

Table 5: DEAR is architecture invariant and can run with any type of processor. Notation taken from Table 1.

|                                                    | Floyd-W.             | DFS                  | SCC                  | Search               | Sort (Ins.)          | Overall<br>(subset) |
|----------------------------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|---------------------|
| <b>NAR<sup>†◆</sup></b><br>(Triplet-MPNN)          | 61.86% ± 1.57        | 31.20% ± 4.02        | <b>46.84% ± 1.70</b> | <b>93.33% ± 0.58</b> | 63.67% ± 39.97       | 59.18%              |
| <b>DEAR<sup>†</sup></b><br>(ours;<br>Triplet-MPNN) | <b>62.29% ± 2.71</b> | <b>42.73% ± 2.79</b> | 45.12% ± 1.52        | 87.00% ± 5.57        | <b>82.34% ± 9.46</b> | <b>63.90%</b>       |