Feasibility Study and Additional Empirical Observations

In addition to the experiments described in the paper, results of some other experiments are repsented in this writeup. These results and the inferences are very interesting; however, owing to the space limitation, we had to be omit them in the original paper.

Experiment and Objectives

The first set of experiments are aimed at studying the feasibility study of the application of Hypothesis-1 as mentioned in Section-2 of the EDBT'23 paper. The results of these experiments presents the limiting conditions under which Hypothesis-1 operates.

As per of the second experiment, we produce the entire result of pre-fetching in terms F-score, True positives, Flase positives and False negatives. We report the mean and standard deviation of these metrics taken over all the test tasks. These results are summarised in the paper as part of the Empirical observations in Section-3 of the paper; where the proposed algorithm's performance has been validated. We also demonstrate how the findings of feasibility study described by the other set of experiments discussd above are related to the empirical observations pertaining to our proposed algorithm's performance validation.

Experiment Setup

The experiment setup is more or less similar to that described in the paper.

The experiment concerning the feasility study of Hypothesis-1 requires, computation of edge-to-node ratio of each component subgraphs. The mean and standard deviation, considering all node's ratio, are computed for each component subgraph.

The experiment concerning the last experiment requires the True positives, False positives and False negatives to be computed for every component subgraph and for every test tasks. The F-score is computed. The average of these metrics are computed over all the component subgraphs and over all the test tasks. The mean, standard deviation, maximum and minimum values for each metric is reported.

For the other experiment which relates the feasibility study experiments, we need to find if the results (F-score TP, FP, FN) are correlated with the edge-to-node ratio. Hence we sort the edge-to-node ratio such that the subgraph ids are arranged in descending order of edge-to-node ratio. We create the first rank of the subgraphs based on edge-to-node ratio. We also sort each of the metrics (F-score, TP, FP, Fn) in descending order, such that the subgraph ids in each case are arranged in the descending order of the respective metric. We create the second rank of the subgraphs based on each of these metrics.

The rank pairs (edge-to-node vs respective metric) are measured for Kendall-Tau rank correlation. For all the correlation, the subgraophs for which $p\text{-value} < 0.05$ have been considered for finding the mean and sd. The mean Kendall-Tau correlation over all the tasks are computed and reported.

Please Turn Over

Results and Discussions

The result for each experiment and the inferences that we can draw from the observation are presented below.

Dataset Name	F-score	Mean edge-to-node	S.D. edge-to-node	Max edge-to-node	Min node-to-edge
MONTAGE	0.77	19.11	46.46	167.0	1.5
WEBLOG	0.76	21.389	37.95	97.278	1.5
CYBERSHAKE	0.16	2.999	0.3964	3.0527	2.933
GENOME	0.212	7.908	14.264	125.0	1.5

Table 1: Hypothesis Feasibility study

From the results above we can observe that MONTAGE and WELOG, which has high F-score, has higher mean edge-to-node ratio than CYBERSHAKE, which exhibits low F-score. Further, the standard deviation from mean in case of MONTAGE and WEBLOG is very high compared to CYBERSHAKE. The standard deviation is also skewed towards higher edge-to-node ratio as is evident from high edge-to-node ratio with CYBERSHAKE having the least mean and standard deviation. CYVERSHAKE also has the least and disproportionately low F-score.

From the above observation we can infer that for obtaining high F-score (better performance), high mean edge-to-node ratio is required. A high mean edge-to-node ratio means that he graph has more edges than nodes; hence, the files represented by the nodes have more dependency amongst themselves. Hence, we get a dense graph, which leads to better prediction ability for our algorithm. In case of CYBERSHAKE, the average edge-to-node ratio is very small; hence, a sparse graph is obtained leading to poor prediction ability and consequent;y low F-score.

Another interesting observation is that of standard deviation. For CYBERSHAKE which has a low standard deviation means that every subgraph more or less has the same edge-to-node ratio. Hence, all the subgraphs have similar low edge-to-node ratio and sparse. If the standard deviation would have been higher, then there was a possibility that some subgraphs are dense at the cost of other subgraphs. In such cases, the dense subgraphs would have had a chance of presenting a good solution with a higher F-score. It may be observed that in the other three datasets, the standard deviation, like mean is highly correlated to the F-score.

The results of the GENOME dataset presents a confirmatory result, to the results of the other 3 datasets discussed above. We observe that while its mean edge-to-node ratio and stdnadard deviation is greater than that of CYBERSHAKE, the values are still substantially lower than the other two datasets- MONTAGE and WEBLOG. The f-score is hence lower than MONTAGE and WEBLOG, but higher than CYBERSHAKE.

In summary, better performance is observed when edge-to-node ratio is high leading to a dense graph. Hence, Hypothesis-1 and threby the solution algorithm is valid for cases where a dense graph is available. This is consistent with the necessity for having high information density that is associated with the good performance of any prediction system.

Dataset Name	Mean of prediction metrics				
	F-score	True Positive	True Negative	False Positive	False Negative
MONTAGE	0.77	0.9900	Not computed since not necessary for F-score	0.5698	0.5745
WEBLOG	0.76	0.8952		0.4948	0.0451
CYBERSHAKE	0.16	0.1691		0.1795	1.75 e-05
GENOME	0.212	0.232		0.194	6.12 e-05

Table 2: Prediction performance in more granular terms

We have described in the paper that MONTAGE exhibits the highest F-score. This high F-score is the result of a very high True positive (hit ratio) of nearly 100%. However, the False postive (file under usage) of more than 50% and True negative of around 57% leaves much room for improvement in future. WEBLOG has been used as a confirmation test dataset and it follows MONTAGE results very closely, confirming our point. CYBERSHAKE was taken as a contrarion example. The sparsity leads to poor prediction and hence, leads to poor performance across all the 4 metrics. Hence, where Hypothesis-1 is applicable, our solution algorithm produces good prediction results.

There is less scope for improvement in True positive, where in case of MONTAGE almost theoretical maximum limit has been reached. Among False positives and Flase negatives, Flase positive is a more pressing concern that should be addressed in future.

The results of GENOME presents a confirmation to the results of the other three datasets and to the inference as discussed above.

Raw results extracted from Colab console output

MONTAGE

```
Mean of edge-to-node ratio 19.11497005988024
Standard Deviation of edge-node ratio:  46.46450605641302
Maximum node-to-edge ratio: 167.0
Minimum node-to-edge ratio: 1.5

Loading the node-edge ratio
Number of files in all subgraph: 872
Mean F1 score: 0.7705625401102402 Standard deviation: 8.122193807987387
Maximum F1 score: 0.7785714285714286 Minimum F1 score: 0.0
Mean TP score: 0.9900900900900901 Standard deviation: 10.435991084218312
Maximum TP score: 1.0 Minimum TP score: 0.0
Mean FP score: 0.5698330440531515 Standard deviation: 0.0047082403505182
Maximum FP score: 0.5745412844036697 Minimum FP score: 0.5688073394495413
Mean FN score: 0.0003985866600545501 Standard deviation:
0.000748202330771138
Maximum FN score: 0.38188073394495414 Minimum FN score: 0.0
Mean Tau correlation between F score and edge-to-node ratio: 1.0
Mean Tau correlation between TP and edge-to-node ratio: 1.0
Mean Tau correlation between FP and edge-to-node ratio: 1.0
Mean Tau correlation between FN and edge-to-node ratio: 1.0
```

CYBERSHAKE

```
Mean of edge-to-node ratio 2.9991648239461104
Standard Deviation of edge-node ratio: 0.03964579515979669
Maximum node-to-edge ratio: 3.05265357292102
Minimum node-to-edge ratio: 2.9333333333333333

Loading the node-edge ratio
Number of files in all subgraph: 39982
Mean F1 score: 0.1564818680660561 Standard deviation: 36.3366082215165
Maximum F1 score: 0.9176497590084922 Minimum F1 score: 0.0
Mean TP score: 0.16906906906906943 Standard deviation: 39.333931384467796
Maximum TP score: 1.0 Minimum TP score: 0.0
Mean FP score: 0.17954311422117258 Standard deviation: 6.234787631739791e-05
Maximum FP score: 0.1795558001100495 Minimum FP score: 0.17948076634485519
Mean FN score: 3.7503362999833436e-05 Standard deviation:
3.7503362999833436e-05
Maximum FN score: 7.503376519433746e-05 Minimum FN score: 0.0
Mean Tau correlation between F score and edge-to-node ratio: 1.0
Mean Tau correlation between TP and edge-to-node ratio: 1.0
Mean Tau correlation between FP and edge-to-node ratio: 1.0
Mean Tau correlation between FN and edge-to-node ratio: 1.0
```

GENOME (Not included in main paper)

```
Mean of edge-to-node ratio 7.908694283641937
Standard Deviation of edge-node ratio: 14.264365196936927
Maximum node-to-edge ratio: 125.0
Minimum node-to-edge ratio: 1.5
Loading the node-edge ratio Number of files in all subgraph: 16338 Mean F1
score: 0.21185512141924734 Standard deviation: 20.23257787758947 Maximum F1
score: 0.9114390114641154 Minimum F1 score: 0.0 Mean TP score:
0.23244026068066617 Standard deviation: 22.19849888264995 Maximum TP score:
1.0 Minimum TP score: 0.0 Mean FP score: 0.19442619166841915 Standard
deviation: 2.8453943038797558e-05 Maximum FP score: 0.19445464561145795
Minimum FP score: 0.19433223160729587 Mean FN score: 6.120700208103753e-05
Standard deviation: 5.421010862427522e-19 Maximum FN score:
0.01524054351817848 Minimum FN score: 0.0 Mean Tau correlation between F
score and edge-to-node ratio: 0.9999999999999999 Mean Tau correlation
between TP and edge-to-node ratio: 0.9999999999999999 Mean Tau correlation
between FP and edge-to-node ratio: 0.9999999999999999 Mean Tau correlation
between FN and edge-to-node ratio: 0.9999999999999999
```

WEBLOG

```
Mean of edge-to-node ratio 21.388972431077697
Standard Deviation of edge-node ratio: 37.952297392997814
Maximum node-to-edge ratio: 97.2781954887218
Minimum node-to-edge ratio: 1.5

Loading the node-edge ratio
Number of files in all subgraph: 288
Mean F1 score: 0.7661756905521191 Standard deviation: 0.09191418092137596
Maximum F1 score: 0.8311688311688312 Minimum F1 score: 0.701182549935407
Mean TP score: 0.8951612903225806 Standard deviation: 0.14826432508750192
Maximum TP score: 1.0 Minimum TP score: 0.7903225806451613
```

```
Mean FP score: 0.4947916666666667 Standard deviation: 0.08854166666666669
Maximum FP score: 0.5833333333333334 Minimum FP score: 0.40625
Mean FN score: 0.04513888888888889 Standard deviation: 0.04513888888888889
Maximum FN score: 0.09027777777777778 Minimum FN score: 0.0
Mean Tau correlation between F score and edge-to-node ratio:
0.9999999999999999
Mean Tau correlation between TP and edge-to-node ratio: 0.9999999999999999
Mean Tau correlation between FP and edge-to-node ratio: 0.9999999999999999
Mean Tau correlation between FN and edge-to-node ratio: 0.9999999999999999
```