

Online Appendix: Structural Patterns and Generative Models of Real-world Hypergraphs

In this online appendix, we provide a complete set of numerical data and figures to supplement the paper "Structural Patterns and Generative Models of Real-world Hypergraphs". The results include patterns, graph measures, and results of several statistical tests on both real and synthetic datasets.

A APPENDIX: DECOMPOSITION

We provide further information on our decomposition method.

A.1 Time complexity of decomposition

THEOREM 1. (*TIME COMPLEXITY OF DECOMPOSITION*). *The time complexity of Decomposition on hypergraph $G = (V, E)$ is: $\sum_{e \in E} \sum_{k=1}^{|e|} O(\binom{|e|}{k}) + O(\binom{\binom{|e|}{k}}{2})$*

PROOF. Consider a hyperedge $e \in E$, e can be decomposed up to the $|e|$ -level. For each decomposition level k , in the k -level decomposed graph $G(k)$: e is decomposed into a clique of $\binom{|e|}{k}$ distinct nodes and this clique has $\binom{\binom{|e|}{k}}{2}$ distinct edges. Therefore, it takes the time $O(\binom{|e|}{k}) + O(\binom{\binom{|e|}{k}}{2})$ to write these nodes and edges into $V_{(k)}$ and $E_{(k)}$, respectively.

Summing over all decomposition levels for e and summing over all hyperedges of G , the total time complexity for the decomposition process is:

$$\sum_{e \in E} \sum_{k=1}^{|e|} O(\binom{|e|}{k}) + O(\binom{\binom{|e|}{k}}{2}) \quad (1)$$

The proof is completed here. \square

Note that if the hyperedge sizes in G is upper-bounded by \bar{s} , the complexity to decompose each hyperedge e is upper-bounded by $\sum_{k=1}^{\bar{s}} O(\binom{\bar{s}}{k}) + O(\binom{\binom{\bar{s}}{k}}{2})$, or $O(1)$. Therefore, the time complexity of Decomposition on G is $O(|E|)$.

A.2 Distinguishing between n-level decomposition and n-order expansion

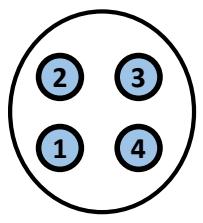
We clarify the difference between the *n-level decomposed graph* in our decomposition method and the *n-projected graph* in "How Much and When Do We Need Higher-order Information in Hypergraphs? A Case Study on Hyperedge Prediction".

¹

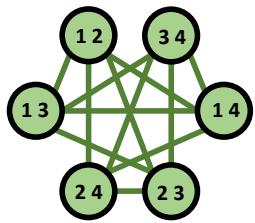
- In n-projected graph, 2 nodes, 2 subsets of nodes in the original hypergraph, are connected if and only if their union is a set of n nodes and is contained in an original hyperedge.
- In n-level decomposed graph, we relax the condition on the union, thus revealing all possible interactions between size- $(n - 1)$ subsets.

For example, for a hyperedge $e = \{1, 2, 3, 4\}$, the difference between the projection of e in 2-projected graph and decomposition of e in 2-level decomposed graph is illustrated in Fig. 1.

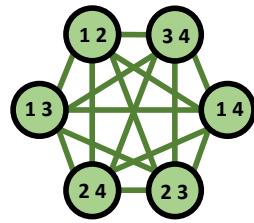
¹Se-eun Yoon, Hyungseok Song, Kijung Shin, and Yung Yi. 2020. *How Much and When Do We Need Higher-order Information in Hypergraphs? A Case Study on Hyperedge Prediction*. In WWW.



Hyperedge e



2-projected graph



2-level decomposed graph

Fig. 1. Compare 2-projected graph and 2-level decomposed graph.

B APPENDIX: SUPPLEMENTARY INFORMATION

We provide some additional information on the datasets and our generators.

The cumulative distribution of hyperedge sizes in each dataset are plotted in Fig. 2. The numbers of nodes edges in the decomposed graphs, up to the level in which the graph maintains a giant connected component of the datasets are provided in Table 1 and 2.

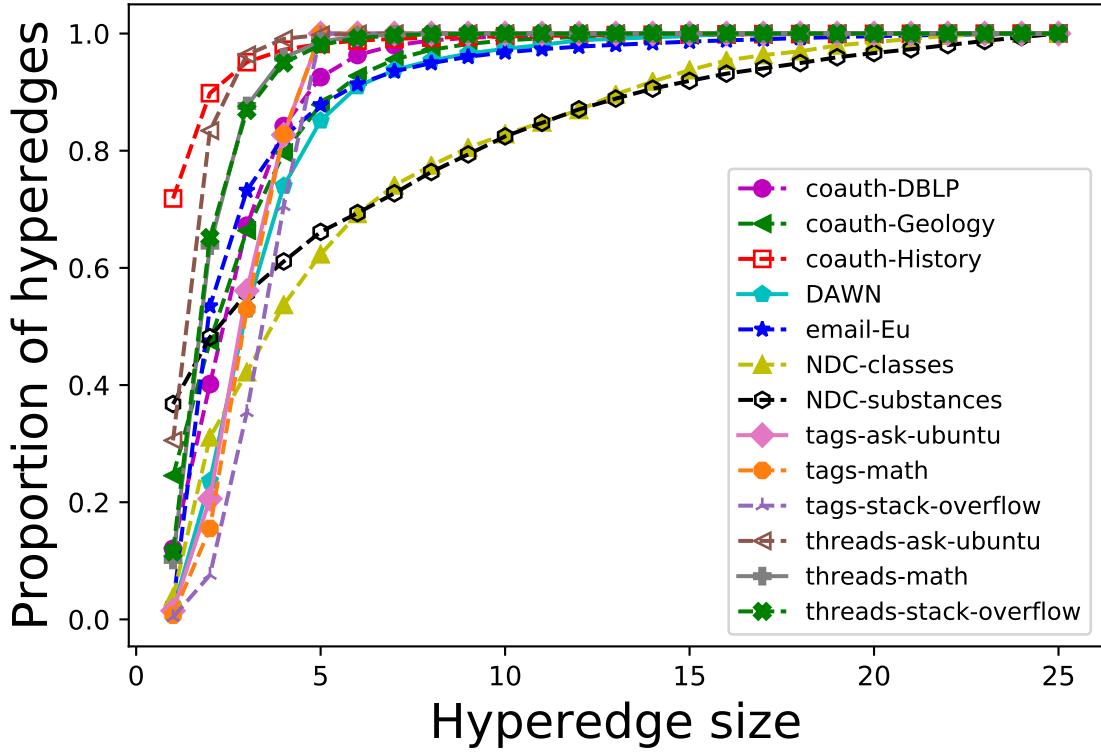


Fig. 2. Cumulative distribution of hyperedge sizes of the hypergraph datasets.

B.1 Size of the largest connected component

The sizes of the largest connected component, in terms of the proportion of the total number of nodes, are listed in Table 3.

Degree Distributions of Decomposed Graphs

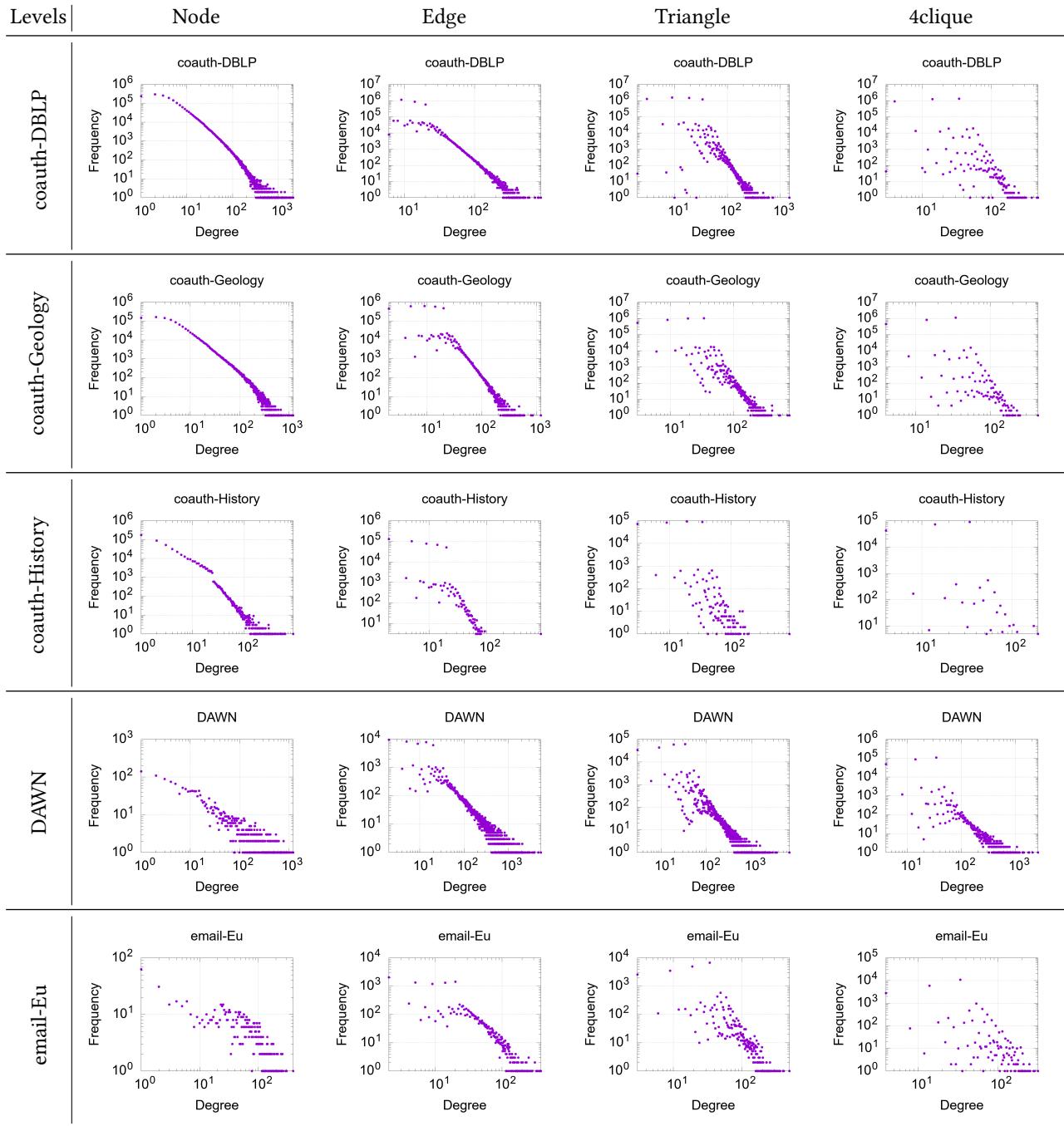


Fig. 3. Degree distributions of decomposed graphs at all decomposition levels.

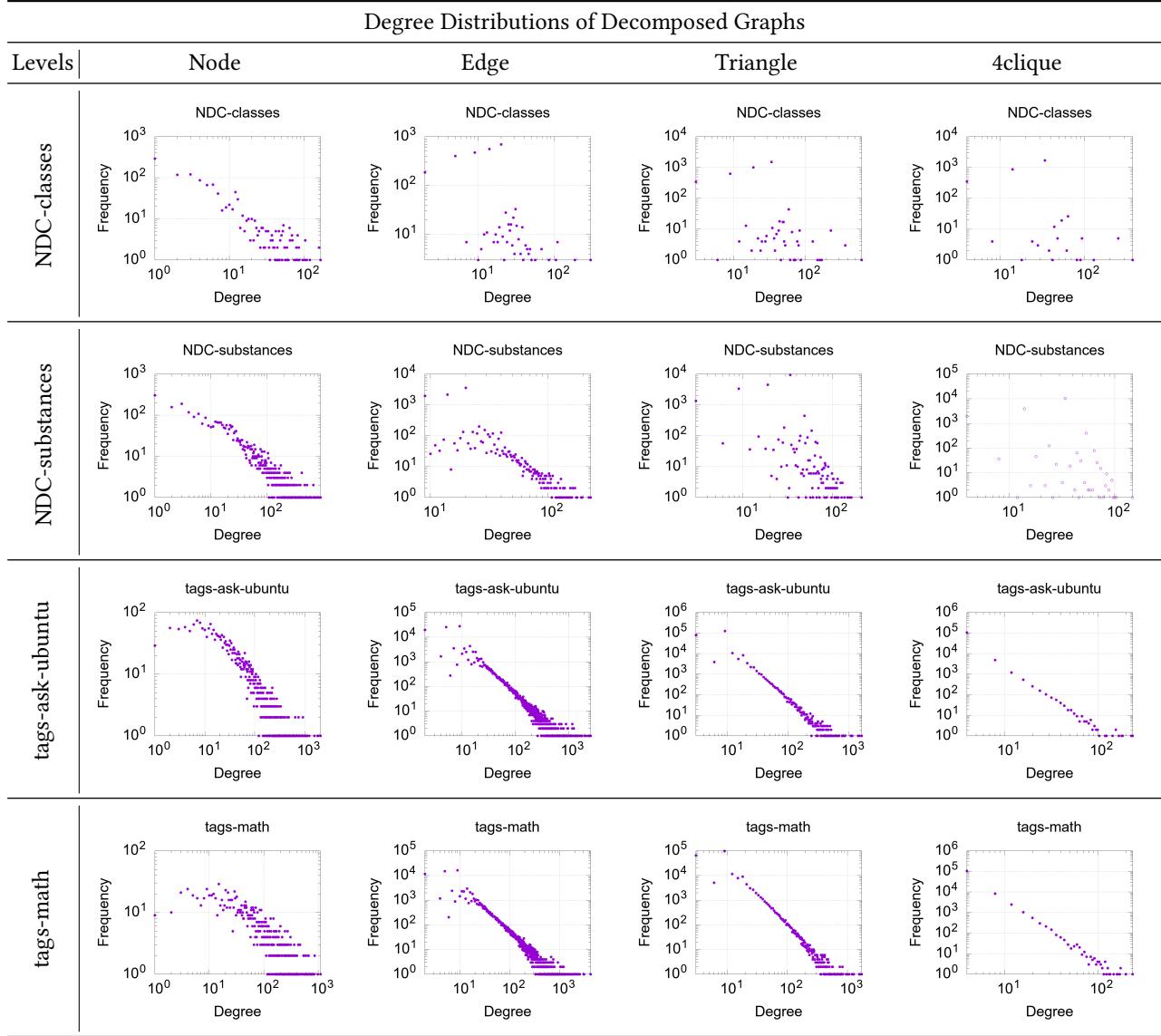


Fig. 4. Degree distributions of decomposed graphs at all decomposition levels (cont).

Degree Distributions of Decomposed Graphs

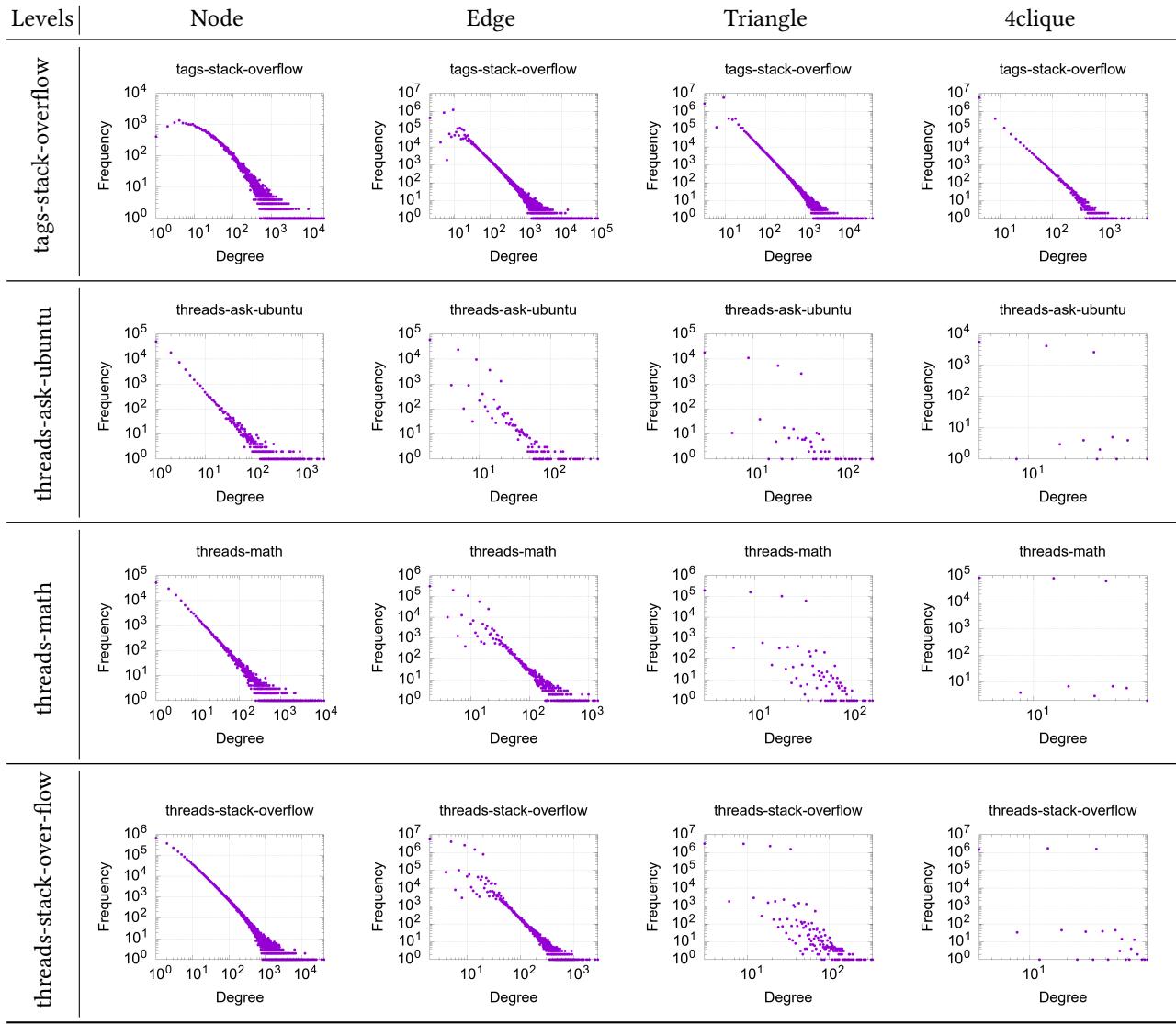


Fig. 5. Degree distributions of decomposed graphs at all decomposition levels (cont).

Table 6. Numerical properties of edge or higher-level decomposed graphs of real-world datasets. As the decomposition level increases, fewer datasets retain giant connected components, and the properties of such datasets are reported in the table. In them, small diameters and high clustering coefficients are observed.

Measure	Nodes	Connect. Comp.	Eff. Diam.	Clust. Coeff.
Edge-level decomposed graphs				
coauth-DBLP	5,906,196	0.57	18.6	0.93
coauth-Geology	3,175,868	0.50	16.4	0.94
DAWN	72,288	0.98	3.9	0.72
email-Eu	13,499	0.98	5.71	0.81
NDC-classes	2,658	0.62	6.6	0.94
NDC-substances	12,882	0.812	9.4	0.89
tags-ask-ubuntu	126,518	0.98	4.5	0.75
tags-math	88,367	0.99	3.9	0.71
tags-stack-overflow	4,083,464	0.99	3.9	0.78
threads-math	782,102	0.61	7.4	0.94
threads-stack-overflow	15,108,684	0.32	12	0.97
Triangle-level decomposed graphs				
DAWN	257,416	0.91	5.3	0.87
email-Eu	24,993	0.86	10.3	0.89
NDC-substances	20,729	0.36	9.4	0.96
tags-ask-ubuntu	248,596	0.79	7.8	0.89
tags-math	222,853	0.91	6.7	0.85
tags-stack-overflow	10,725,751	0.92	6.5	0.88
4clique-level decomposed graphs				
DAWN	284,755	0.52	8.1	0.89
email-Eu	24,772	0.41	15.3	0.89
tags-ask-ubuntu	145,676	0.22	17.1	0.74
tags-math	156,129	0.35	14.8	0.71
tags-stack-overflow	7,887,748	0.42	13	0.76

Distributions of Singular Values of Decomposed Graphs

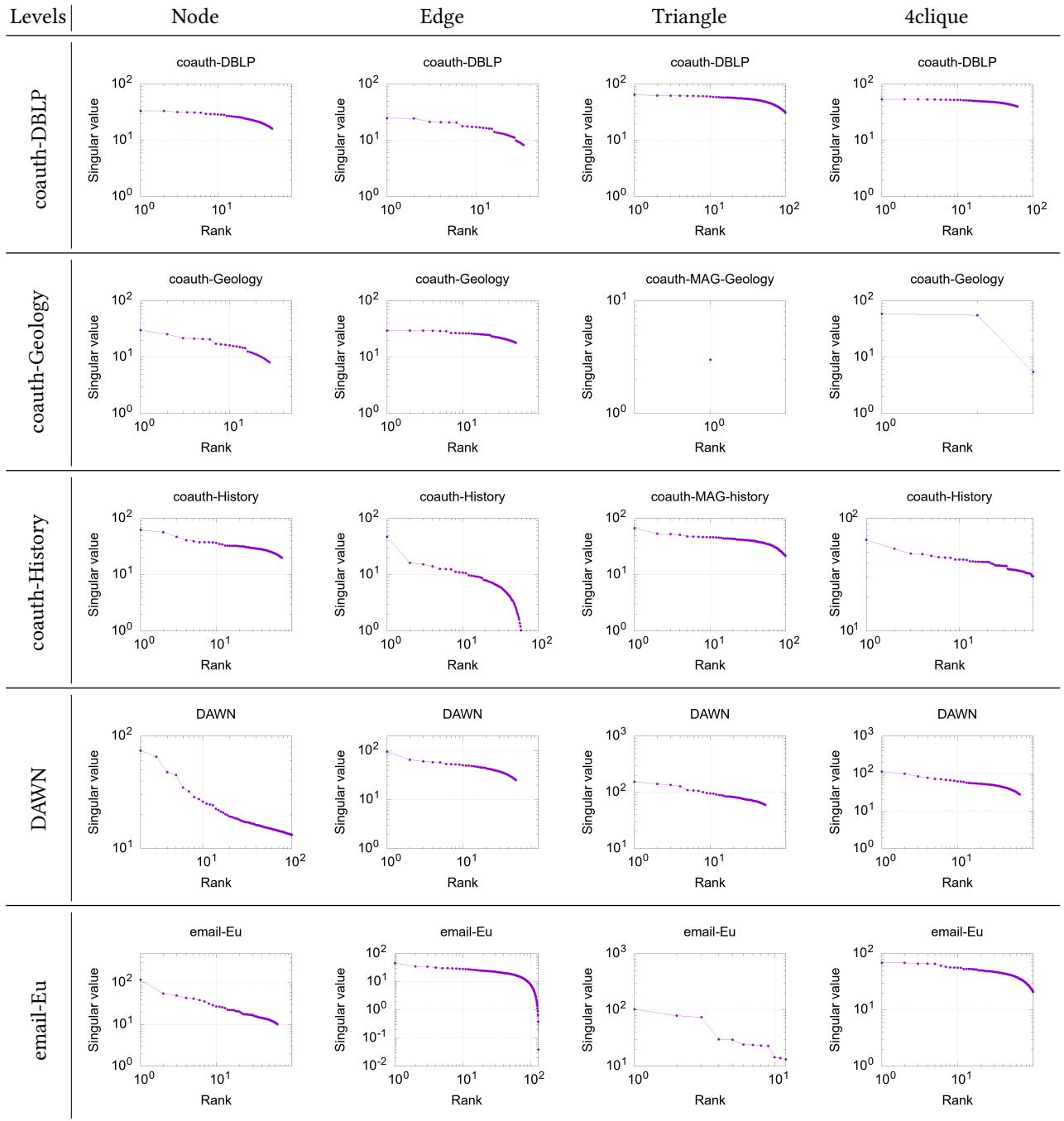


Fig. 6. Distributions of singular values of decomposed graphs at all decomposition levels.

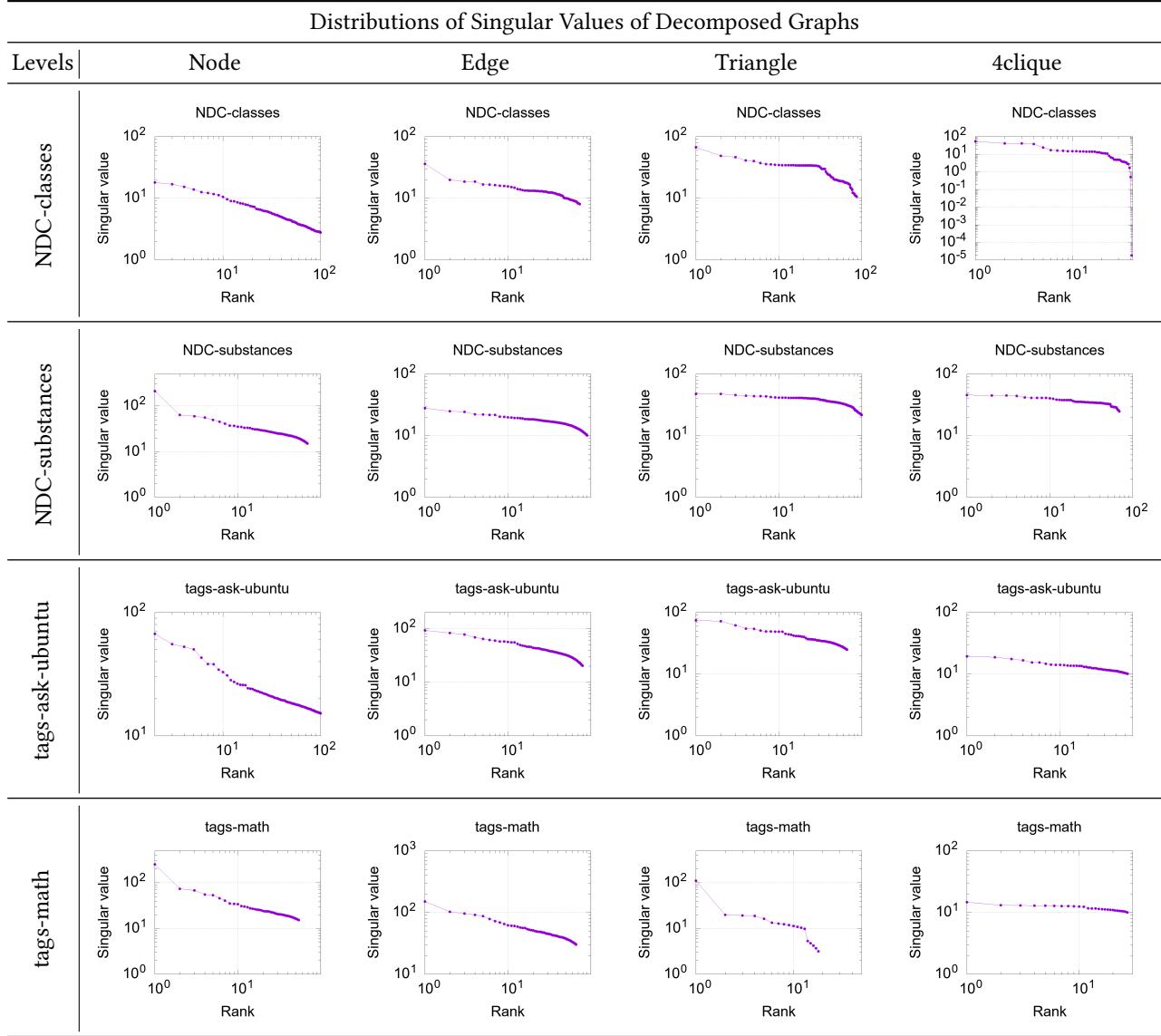


Fig. 7. Distributions of singular values of decomposed graphs at all decomposition levels (cont).

Distributions of Singular Values of Decomposed Graphs

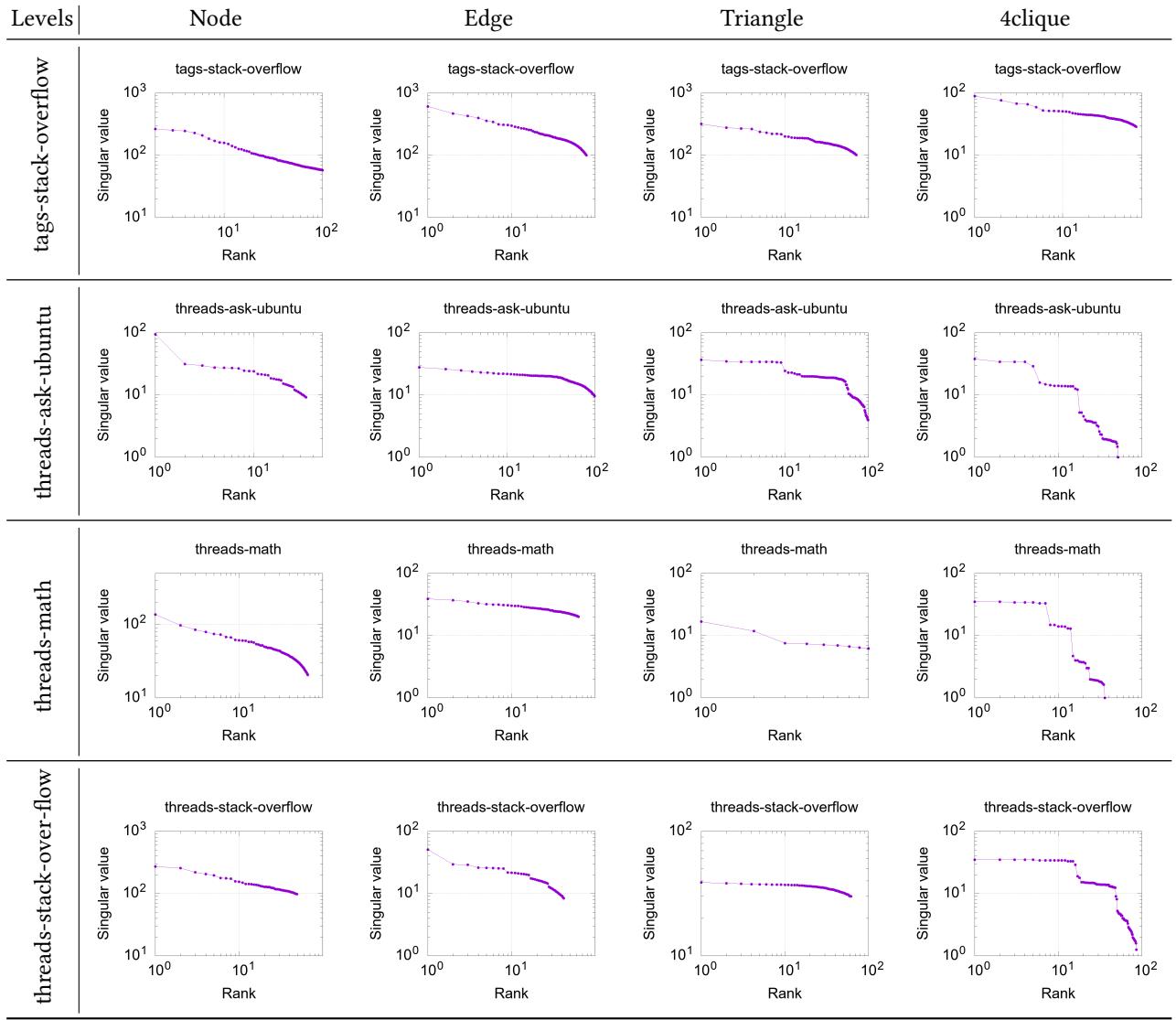


Fig. 8. Distributions of singular values of decomposed graphs at all decomposition levels (cont).

C APPENDIX: RESULTS OF THE GENERATORS

We provide patterns and graph measurements on the 4 considered synthetic datasets.

C.1 Size of the largest connected component

The sizes of the largest connected components are reported in terms of the proportion of the total number of nodes in Table 8. The red entries correspond to the case where the pattern ‘giant connected component’ does not exist.

Table 8. Sizes of the largest connected components of real and synthetic datasets at all 4 levels.

Dataset	Level	Real Data	HYPERPA (Proposed)	Naive PA	Subset Sampling
DAWN	Node	0.89	0.996	0.73	0.999
	Edge	0.98	0.98	0.95	0.95
	Triangle	0.91	0.89	0.08	0.79
	4clique	0.52	0.81	0.01	0.22
email-Eu	Node	0.98	0.995	0.997	0.988
	Edge	0.98	0.86	0.935	0.8
	Triangle	0.86	0.86	0.54	0.5
	4clique	0.41	0.76	0.03	0.04
tags-ask-ubuntu	Node	0.99	0.99	0.99	0.99
	Edge	0.98	0.92	0.98	0.95
	Triangle	0.79	0.81	0.74	0.55
	4clique	0.21	0.39	0.11	0.002
tags-math	Node	0.99	0.997	0.997	0.996
	Edge	0.99	0.98	0.993	0.97
	Triangle	0.91	0.81	0.77	0.55
	4clique	0.35	0.28	0.12	0.02

C.2 Degree distribution

Plots of degree distributions at all 4 decomposition levels of the synthetic datasets generated by the three models are given in Figures 9, 10, 11 and 12. In addition, D-statistics between the degree distributions of the corresponding real and generated datasets are listed in Table 9. In each row, the smallest value, implying the closest distance between the synthetic and real distributions, is in bold.

Table 9. D-statistics between the degree distributions of real dataset and synthetic datasets generated by the 3 models. We generated each dataset 5 times and report the average. Values higher than 0.2 are in red.

Dataset	Level	HYPERP A	Naive	Subset
		(Proposed)	PA	Sampling
DAWN	Node	0.153	0.184	0.132
	Edge	0.135	0.082	0.059
	Triangle	0.117	0.077	0.203
	4clique	0.048	0.041	0.049
email-Eu	Node	0.392	0.282	0.235
	Edge	0.109	0.148	0.126
	Triangle	0.159	0.19	0.178
	4clique	0.128	0.149	0.141
tags-ask-ubuntu	Node	0.065	0.259	0.128
	Edge	0.082	0.232	0.057
	Triangle	0.069	0.428	0.049
	4clique	0.087	0.655	0.029
tags-math	Node	0.2	0.364	0.249
	Edge	0.101	0.216	0.073
	Triangle	0.072	0.365	0.117
	4clique	0.025	0.615	0.077

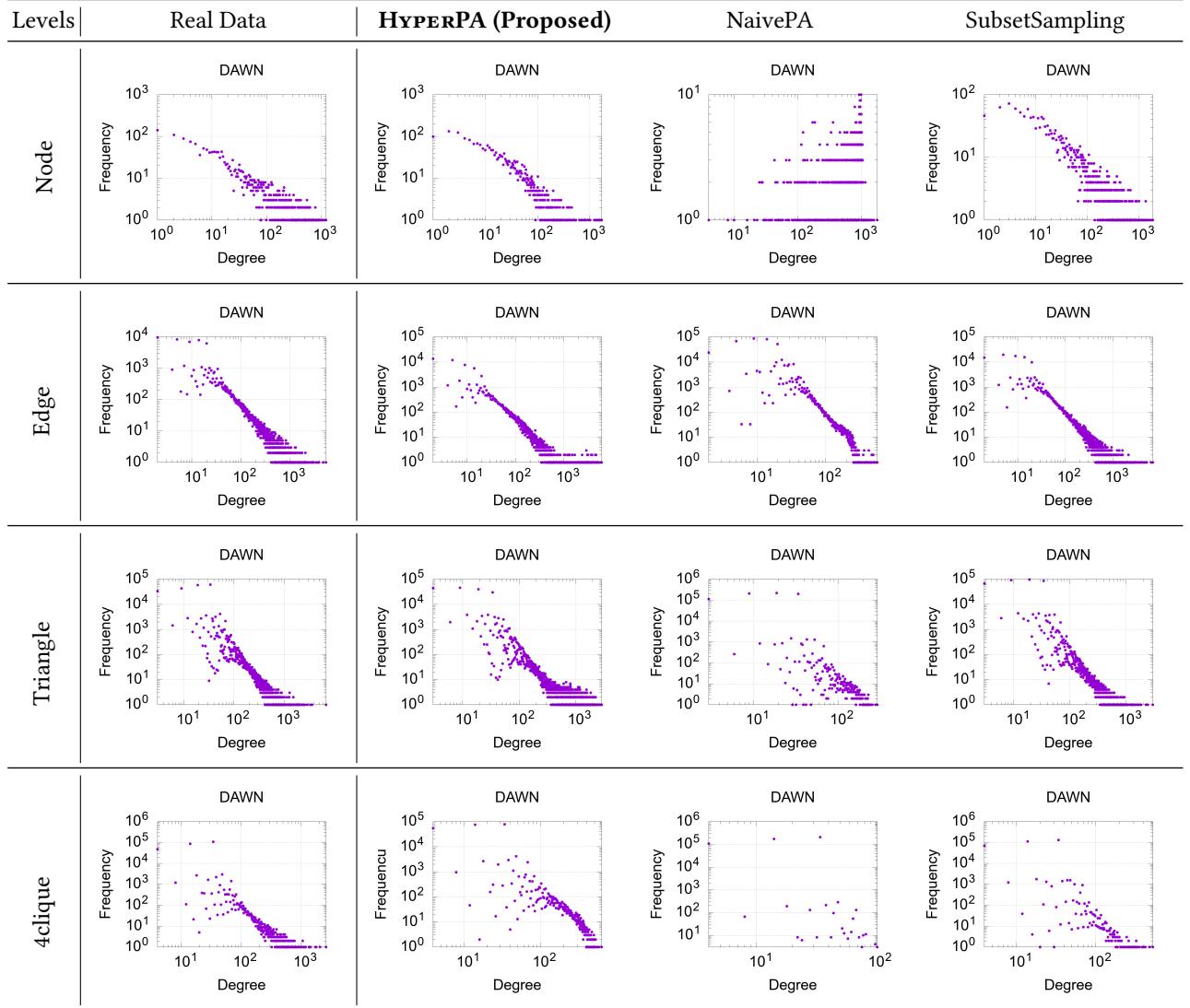


Fig. 9. Comparison of hypergraph generators with respect to degree distributions of decomposed graphs at different decomposition levels. The DAWN dataset was used to learn S , NP and n .

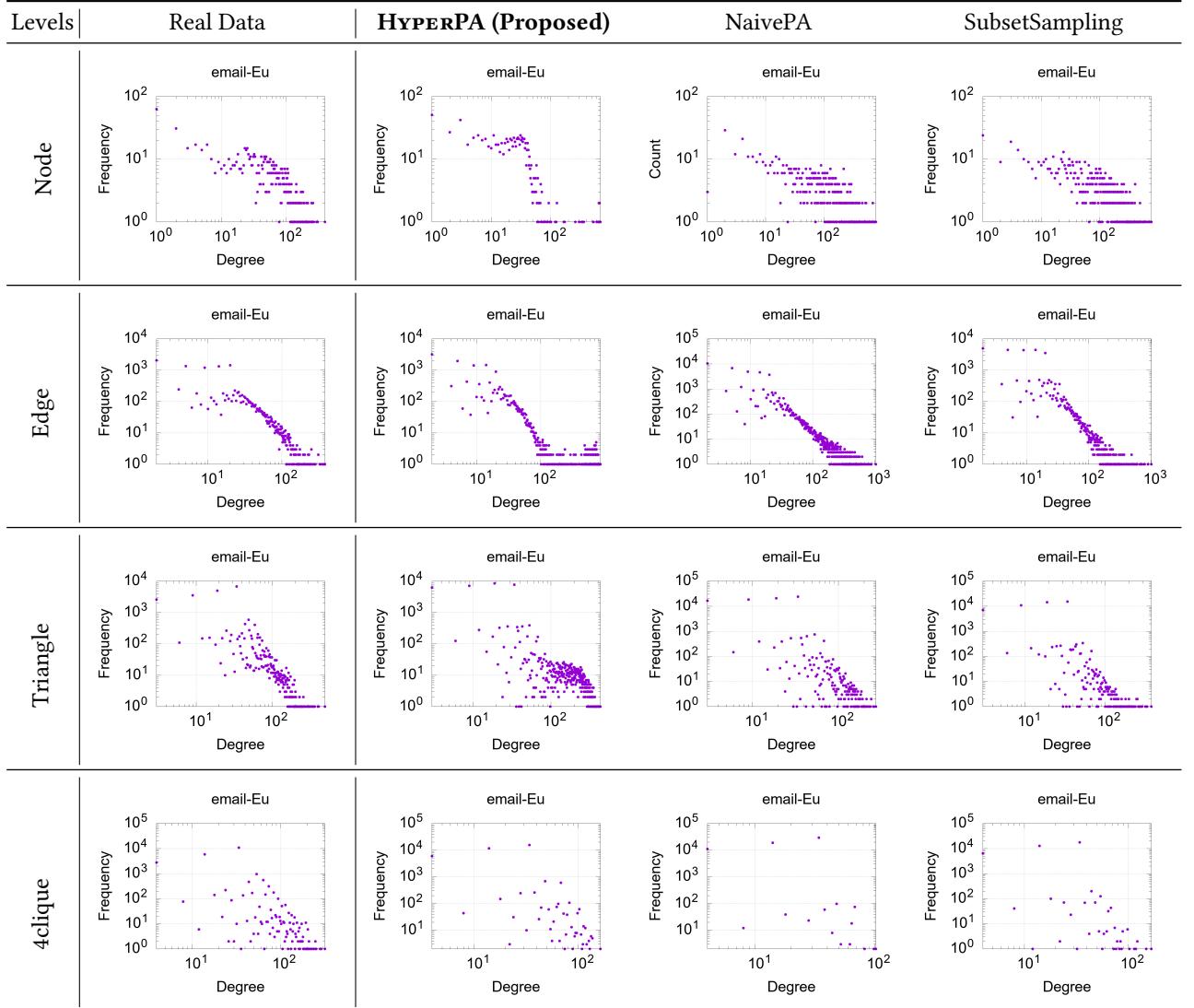


Fig. 10. Comparison of hypergraph generators with respect to degree distributions of decomposed graphs at different decomposition levels. The *email-Eu* dataset was used to learn S , NP and n .

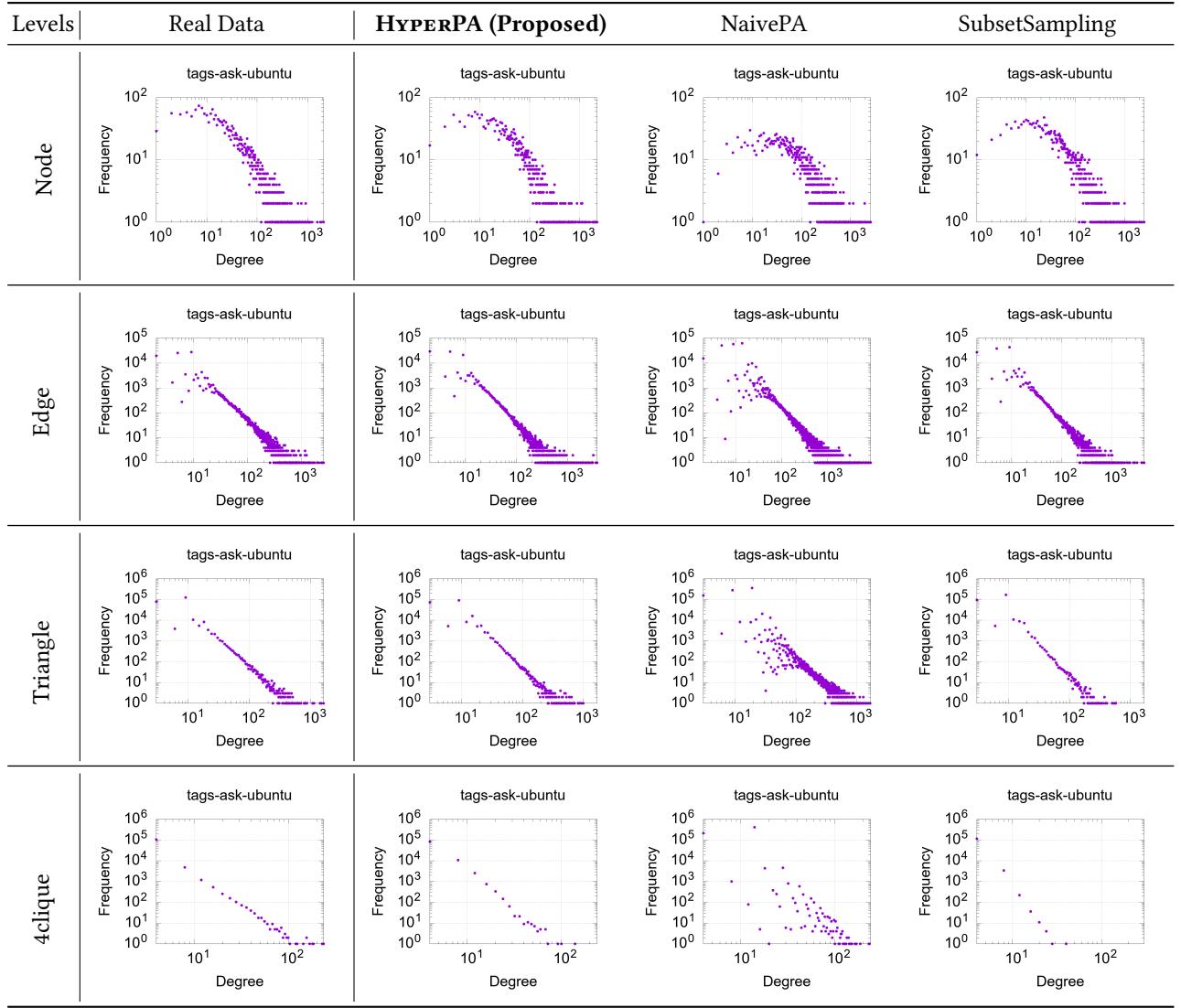


Fig. 11. Comparison of hypergraph generators with respect to degree distributions of decomposed graphs at different decomposition levels. The *tags-ask-ubuntu* dataset was used to learn S , NP and n .

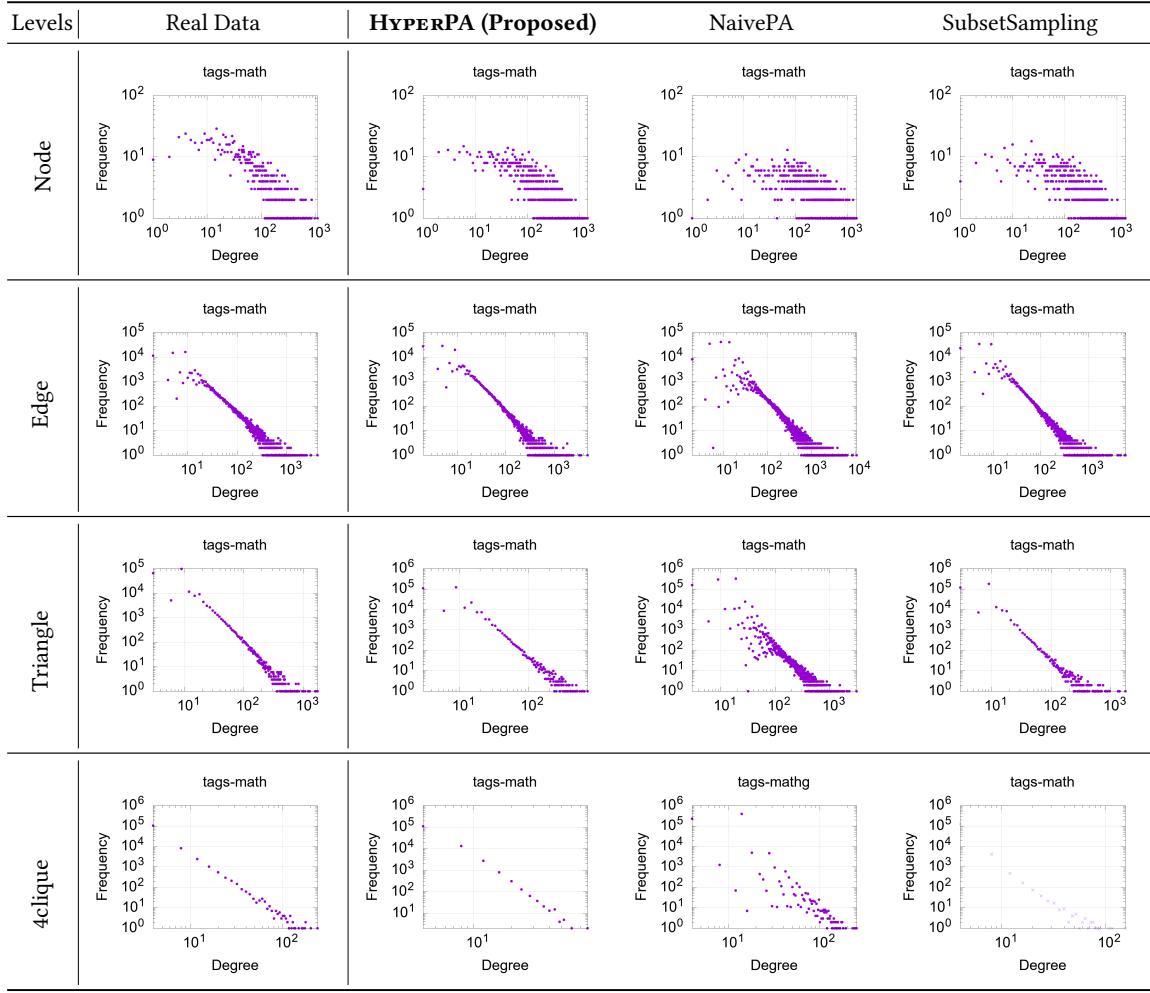


Fig. 12. Degree distributions of decomposed graphs at different decomposition levels of *tags-math*, both real and generated datasets. Those generated by HYPERPA achieve more similarity to the real distribution (confirmed in Table ??).

C.3 Diameter

The effective diameters at all 4 decomposition levels of the generated datasets are reported in Table 10. For each real diameter d , considering the acceptance range of $(\frac{2d}{3}, \frac{4d}{3})$, the values that are not within this range are in red.

Table 10. Effective diameters of real and synthetic datasets in all 4 levels.

Dataset	Level	Real Data	HYPERPA (Proposed)	Naive PA	Subset Sampling
DAWN	Node	2.6	2	1.84	2
	Edge	3.9	3.5	6.8	3.9
	Triangle	5.3	3.9	11.2	5.9
	4clique	8.1	5.5	9.9	8.26
email-Eu	Node	2.8	1.96	1.93	1.96
	Edge	5.7	3.4	4.4	4.8
	Triangle	10.3	3.9	6.4	6.9
	4clique	15.3	6.9	9.15	6.5
tags-ask-ubuntu	Node	2.4	1.95	1.9	1.95
	Edge	4.5	4.4	3.8	4.6
	Triangle	7.8	7	5.77	8.2
	4clique	17.1	15.75	9.1	5.8
tags-math	Node	2.1	1.9	1.88	1.9
	Edge	3.9	4.4	3.76	4.5
	Triangle	6.7	8.2	5.75	7.5
	4clique	14.8	18.9	8.5	8

C.4 Clustering coefficient

The global clustering coefficients at all 4 decomposition levels of the generated datasets are reported in Table 11. For each real clustering coefficient c , considering the acceptance range of $(\frac{2c}{3}, \min(\frac{4c}{3}, 1))$, the values that are not within this range are in red.

Table 11. Clustering coefficients of real and synthetic datasets in all 4 levels.

Dataset	Level	Real Data	HYPERPA (Proposed)	Naive PA	Subset Sampling
DAWN	Node	0.64	0.82	0.37	0.78
	Edge	0.72	0.76	0.82	0.7
	Triangle	0.87	0.77	0.96	0.86
	4clique	0.89	0.85	0.62	0.73
email-Eu	Node	0.49	0.81	0.73	0.63
	Edge	0.81	0.68	0.78	0.71
	Triangle	0.89	0.8	0.85	0.89
	4clique	0.89	0.9	0.6	0.66
tags-ask-ubuntu	Node	0.61	0.6	0.72	0.62
	Edge	0.75	0.71	0.76	0.74
	Triangle	0.89	0.74	0.9	0.83
	4clique	0.74	0.69	0.67	0.34
tags-math	Node	0.63	0.67	0.73	0.65
	Edge	0.71	0.68	0.69	0.7
	Triangle	0.85	0.75	0.9	0.825
	4clique	0.71	0.67	0.68	0.33

C.5 Singular values

Plots of distributions of singular values at 4 levels of the synthetic datasets generated by the three models are given in Figures 13, 14, 15 and 16. In addition, D-statistics between the distributions of singular values of the corresponding real and generated datasets are listed in Table 12. In each row, the smallest value, implying the closest distance between the synthetic and real distributions, is in bold.

Table 12. D-statistic between the singular-value distributions of real dataset and of synthetic datasets generated by the 3 methods. We generated each dataset 5 times and report the average. Values higher than 0.2 are in red.

Dataset	Level	HYPERP A	Naive	Subset
		(Proposed)	PA	Sampling
DAWN	Node	0.2	0.162	0.125
	Edge	0.167	0.227	0.259
	Triangle	0.256	0.21	0.335
	4clique	263	0.37	0.433
email-Eu	Node	0.413	0.185	0.2
	Edge	0.185	0.223	0.216
	Triangle	0.219	0.376	0.497
	4clique	0.408	0.488	0.407
tags-ask-ubuntu	Node	0.226	0.21	0.225
	Edge	0.169	0.397	0.322
	Triangle	0.288	0.373	0.369
	4clique	0.215	0.507	0.521
tags-math	Node	0.228	0.168	0.502
	Edge	0.241	0.348	0.116
	Triangle	0.344	0.491	0.292
	4clique	0.3	0.51	0.369

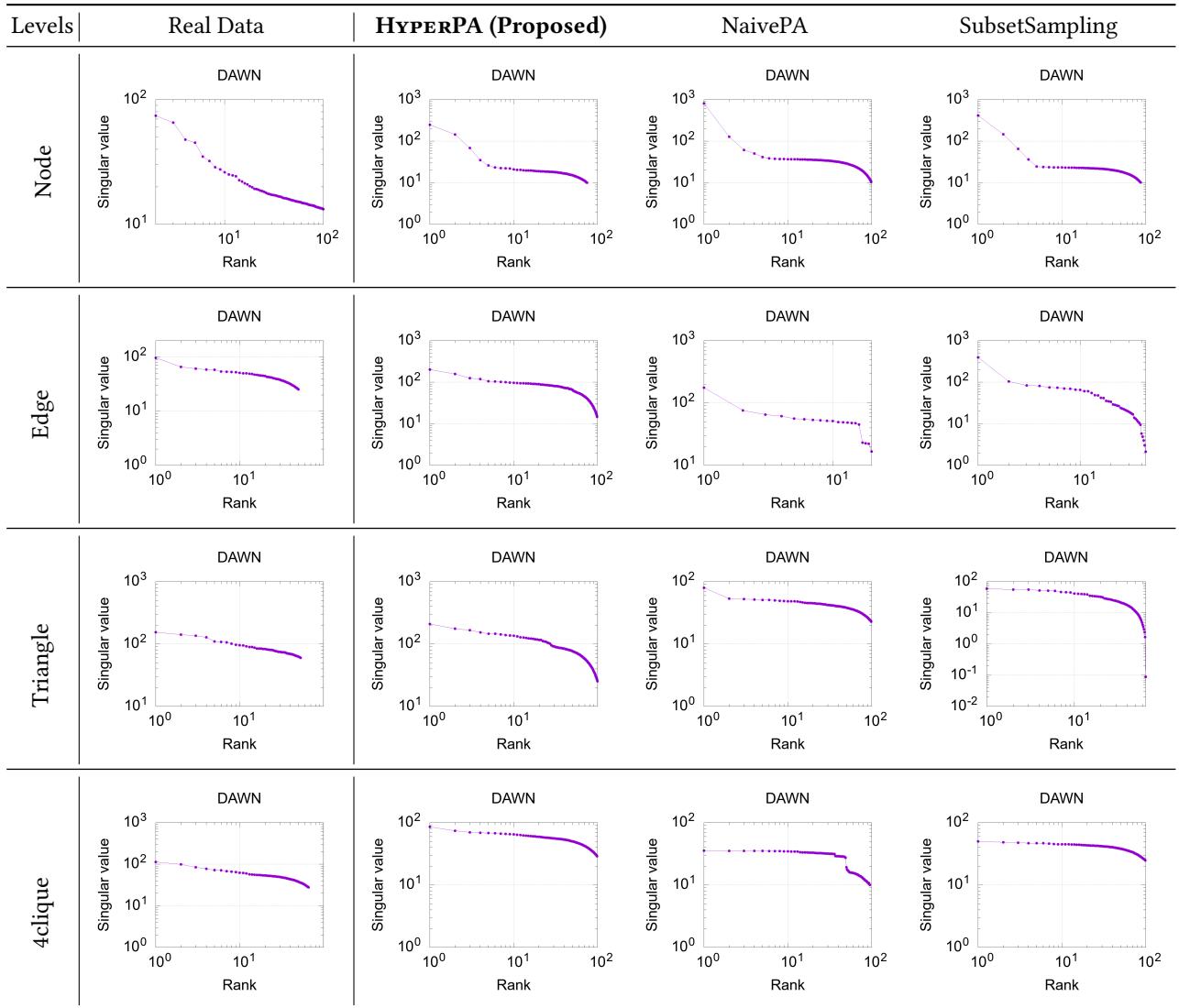


Fig. 13. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The DAWN dataset was used to learn S , NP and n .

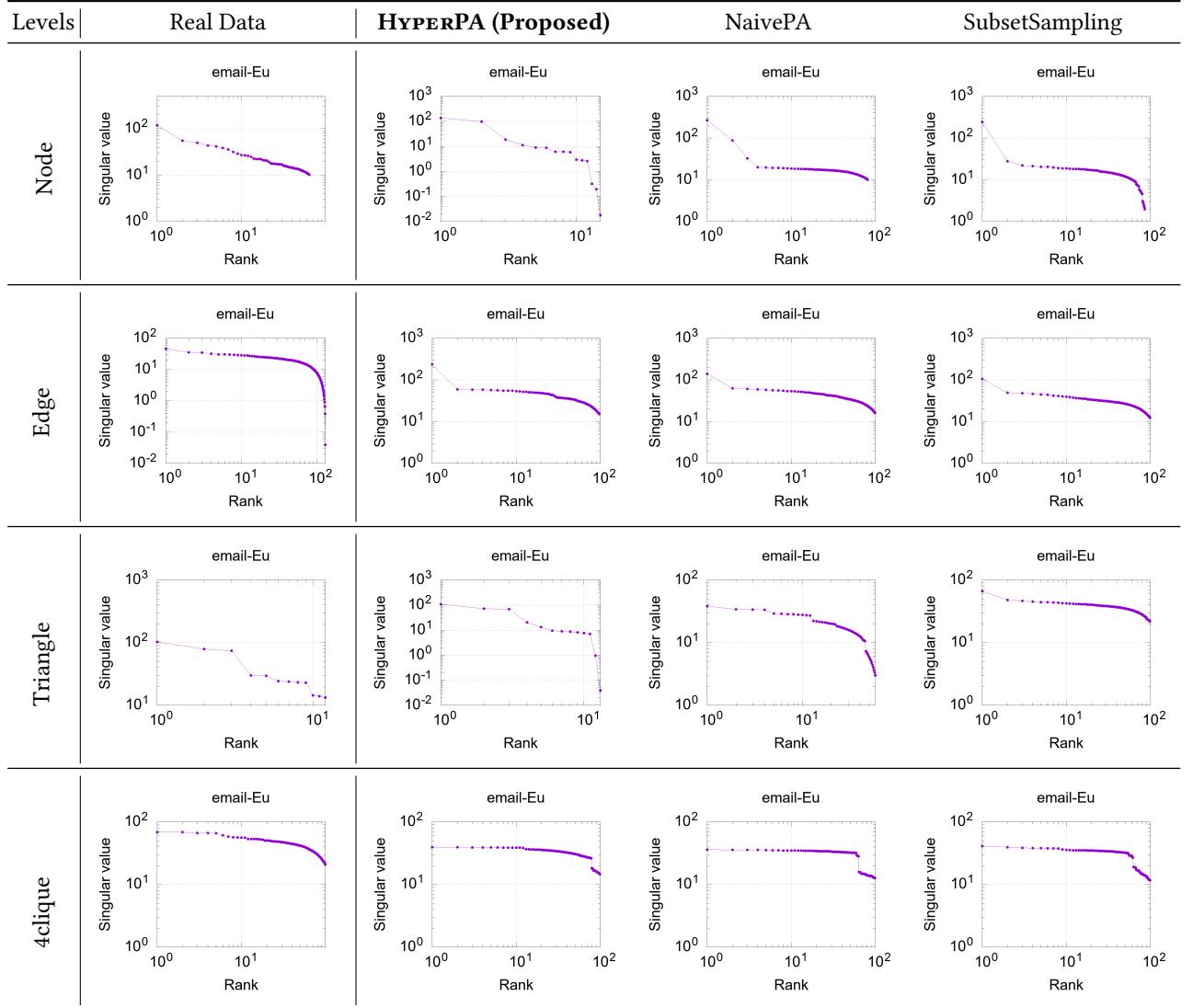


Fig. 14. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *email-Eu* dataset was used to learn S , NP and n .

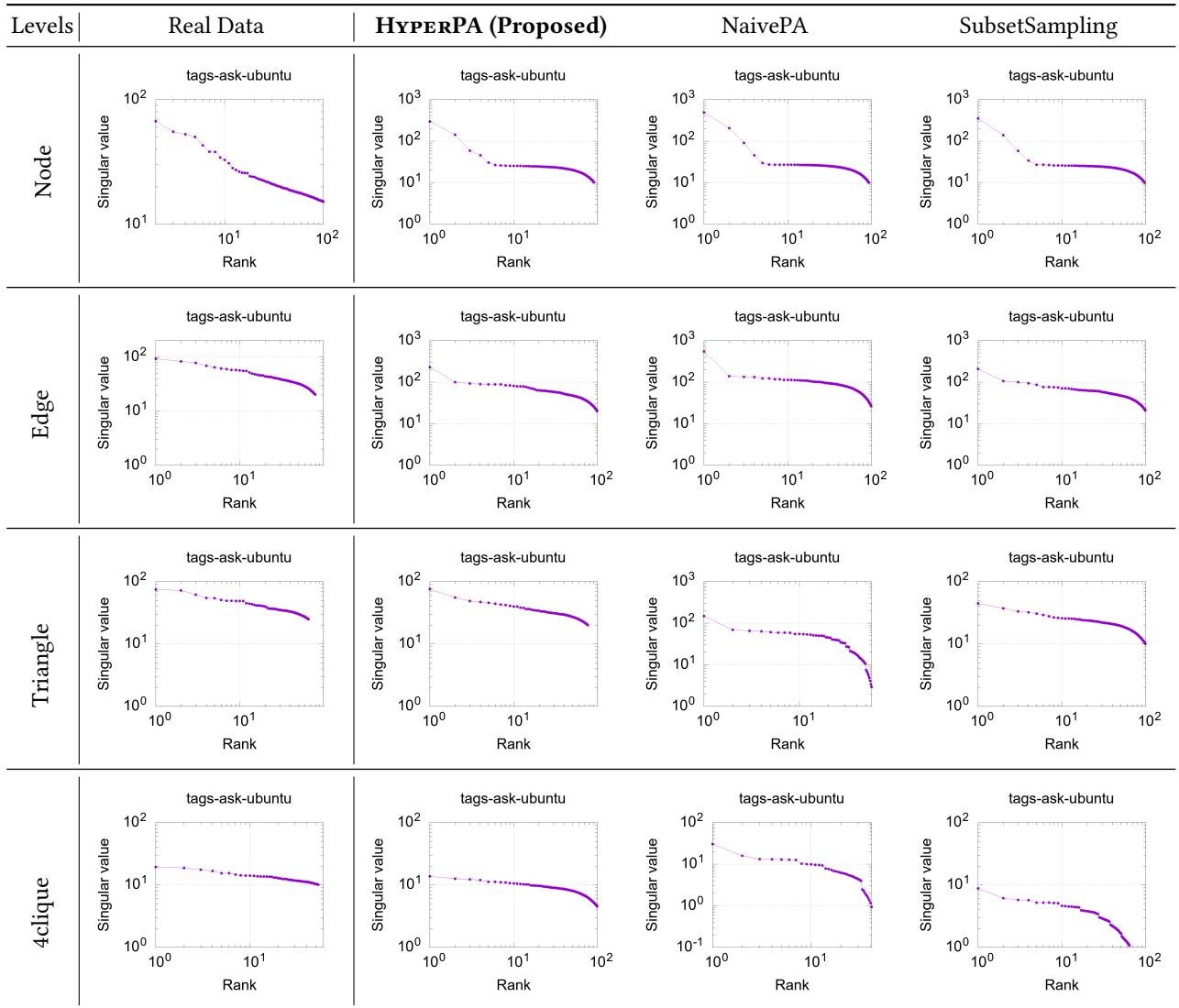


Fig. 15. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *tags-ask-ubuntu* dataset was used to learn S , NP and n .

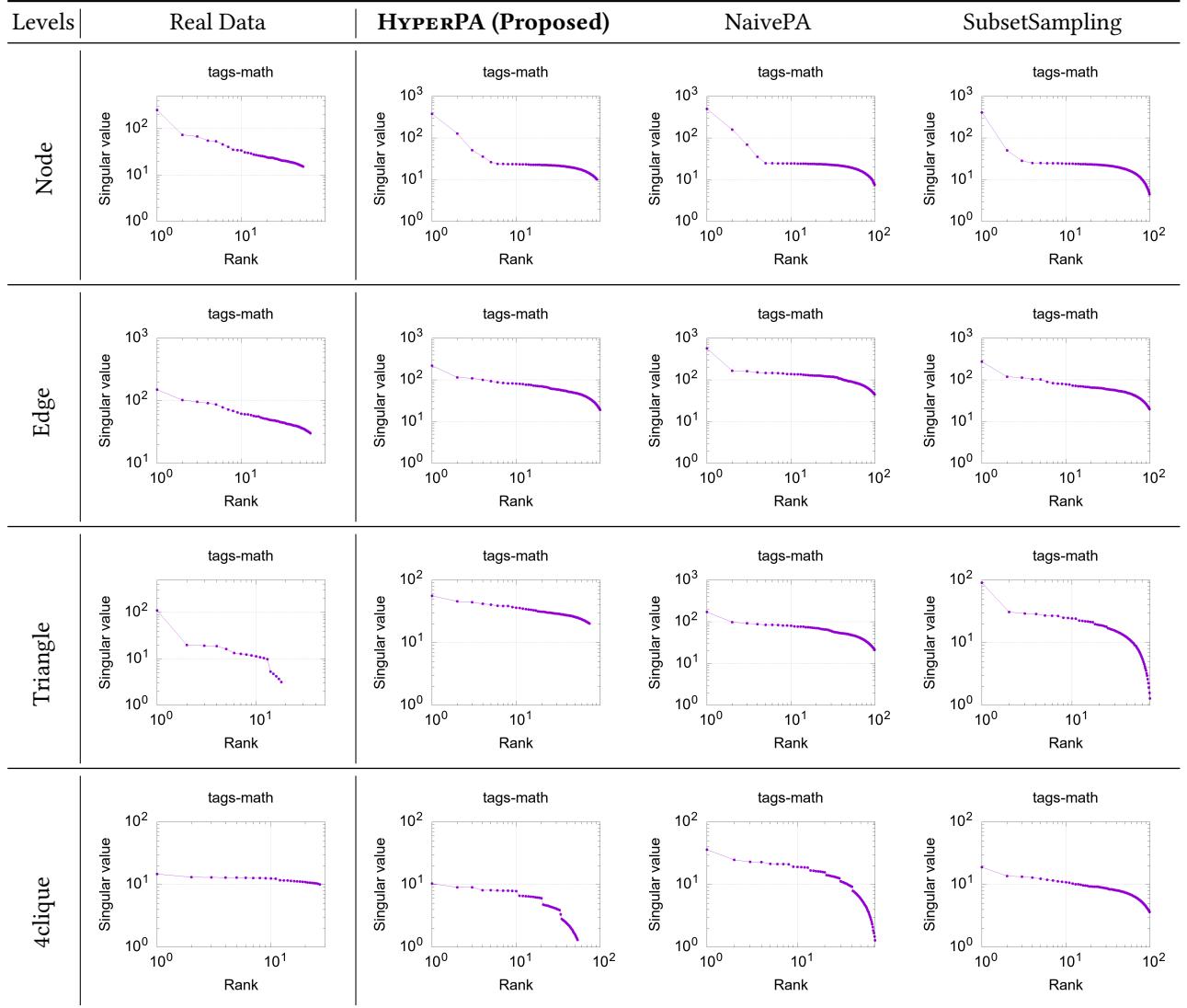


Fig. 16. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *tags-math* dataset was used to learn S , NP and n .