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

MANH TUAN DO [1], SE-EUN YOON [1], BRYAN HOOI[2], KIJUNG SHIN [1]

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" [3]. 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 a hypergraph* $G = (V, E)$ *is :* $O(\sum_{e \in E} \sum_{k=1}^{|e|-1} \binom{\binom{|e|}{k}}{2})$

PROOF. Consider a hyperedge $e \in E$, $e$ can be decomposed up to the $|e| - 1$ 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 $O(\binom{|e|}{k}) + O(\binom{\binom{|e|}{k}}{2}) = O(\binom{\binom{|e|}{k}}{2})$ time to create these nodes and edges forming $V_{(k)}$ and $E_{(k)}$, respectively. Summing over all decomposition levels for $e$ and over all hyperedges of $G$, the total time complexity of the decomposition process is:

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

The proof is completed here. □

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

### 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" [4].

- In the n-projected graph, two nodes (i.e., two size-$(n-1)$ 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 a hyperedge.
- In the n-level decomposed graph, we relax the condition on the size of the union, thus revealing all possible interactions between the size-$(n-1)$ subsets.

[1]Korea Advanced Institute of Science and Technology
[2]National University of Singapore

[3]Manh Tuan Do, Se-eun Yoon, Bryan Hooi, and Kijung Shin. 2020. *Structural Patterns and Generative Models of Real-world Hypergraphs.* In KDD.

[4]Se-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.

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

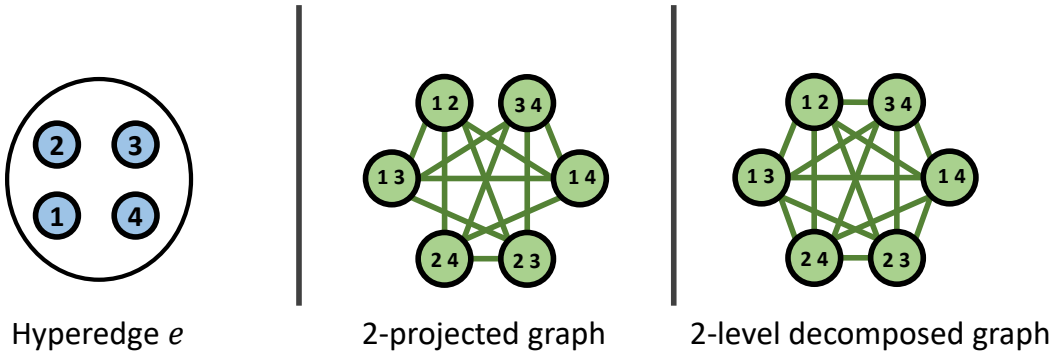


Fig. 1. Compare a 2-projected graph and a 2-level decomposed graph.

## 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 ploted 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.

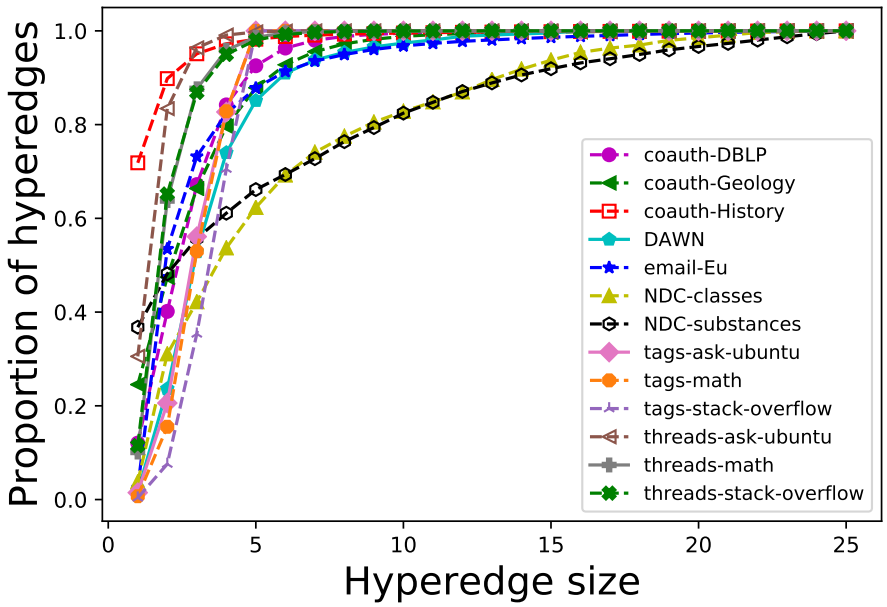


Fig. 2. Cumulative distribution of hyperedge sizes of the hypergraph datasets.

### 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.

Table 1. Number of nodes in the decomposed graphs, up to the level in which the graph retains a giant connected component, of the datasets.

| Dataset | $\lvert V_{(1)}\rvert$ | $\lvert V_{(2)}\rvert$ | $\lvert V_{(3)}\rvert$ | $\lvert V_{(4)}\rvert$ |
|---|---|---|---|---|
| coauth-DBLP | 1,924,991 | 6,316,009 | 6,478,373 | 3,949,887 |
| coauth-Geology | 1,256,385 | 3,371,254 | 3,751,261 | 2,577,988 |
| coauth-History | 1,014,734 | 596,283 | 390,739 | 226,684 |
| DAWN | 2,558 | 81,567 | 276,003 | 284,755 |
| email-Eu | 998 | 18,955 | 27,151 | 24,772 |
| NDC-classes | 1,161 | 2,836 | 3,778 | 3,078 |
| NDC-substances | 5,311 | 13,527 | 21,232 | 17,757 |
| tags-ask-ubuntu | 3,029 | 132,703 | 279,369 | 145,676 |
| tags-math | 1,629 | 91,685 | 249,548 | 156,129 |
| tags-stack-overflow | 49,998 | 4,147,302 | 11,383,997 | 7,887,748 |
| threads-ask-ubuntu | 125,602 | 186,270 | 58,892 | 16,957 |
| threads-math | 176,445 | 1,064,478 | 664,932 | 272,046 |
| threads-stack-overflow | 2,675,995 | 20,077,609 | 12,052,144 | 5,547,010 |

Table 2. Number of edges in the decomposed graphs, up to the level in which the graph retains a giant connected component, of the datasets.

| Dataset | $\lvert E_{(1)}\rvert$ | $\lvert E_{(2)}\rvert$ | $\lvert E_{(3)}\rvert$ | $\lvert E_{(4)}\rvert$ |
|---|---|---|---|---|
| coauth-DBLP | 7,904,336 | 31,284,160 | 50,887,503 | 35,299,764 |
| coauth-Geology | 5,120,762 | 18,987,747 | 35,384,178 | 26,839,940 |
| coauth-History | 1,156,914 | 1,852,269 | 3,001,774 | 2,183,900 |
| DAWN | 122,963 | 1,682,274 | 4,097,770 | 3,219,360 |
| email-Eu | 29,299 | 155,769 | 393,527 | 360,955 |
| NDC-classes | 6,222 | 20,568 | 45,793 | 38,525 |
| NDC-substances | 88,268 | 116,967 | 268,057 | 231,445 |
| tags-ask-ubuntu | 132,703 | 1,275,135 | 1,256,181 | 254,750 |
| tags-math | 91,685 | 1,217,031 | 1,375,434 | 292,440 |
| tags-stack-overflow | 4,147,302 | 57,815,235 | 71,817,873 | 16,327,590 |
| threads-ask-ubuntu | 187,157 | 227,547 | 175,627 | 85,665 |
| threads-math | 1,089,307 | 2,810,934 | 3,086,411 | 1,770,730 |
| threads-stack-overflow | 20,999,838 | 52,797,462 | 66,240,865 | 41,329,315 |

## B.2 Degree distribution

While the plots for degree distribution are provided at all 4 levels (Figures 3, 4, 5), note that the pattern of 'heavy-tailed degree distribution' is only inspected up to the decomposition level in which the graph retains a 'giant connected component'. In addition, we also provide the loglikehood ratio of fitting empirical distributions to a heavy-tailed distribution versus the exponential distribution (Exp.) in Table 4:

- Power-law (pw) vs. Exp.
- Truncated power-law (trunpw) vs. Exp.
- Lognormal (lgnorm) vs. Exp.

Table 3. Size of the largest connected component, as the proportion of the total number of nodes (including the degree-zero nodes), in each dataset at each decomposition level. The red numbers indicate that the graph no longer retains a giant connected component. In the case of *NDC-classes*, the size of the second largest connected component at triangle and 4clique levels is 0.11 and 0.04. According to the description in Section 4.1 in the main paper, a giant connected component does not exist.

| Level | Node ($k=1$) | Edge ($k=2$) | Triangle ($k=3$) | 4clique ($k=4$) |
|---|---|---|---|---|
| coauth-DBLP | 0.86 | 0.57 | 0.05 | 0.0006 |
| coauth-Geology | 0.72 | 0.5 | 0.06 | 0.0005 |
| coauth-History | 0.22 | 0.002 | 0.002 | 0.001 |
| DAWN | 0.89 | 0.98 | 0.91 | 0.52 |
| email-Eu | 0.98 | 0.98 | 0.86 | 0.41 |
| NDC-classes | 0.54 | 0.62 | 0.27 | 0.19 |
| NDC-substances | 0.58 | 0.82 | 0.36 | 0.02 |
| tags-ask-ubuntu | 0.99 | 0.99 | 0.79 | 0.21 |
| tags-math | 0.99 | 0.99 | 0.91 | 0.35 |
| tags-stack-overflow | 0.99 | 0.99 | 0.92 | 0.42 |
| threads-ask-ubuntu | 0.65 | 0.09 | 0.02 | 0.01 |
| threads-math | 0.86 | 0.61 | 0.03 | 0.0004 |
| threads-stack-overflow | 0.86 | 0.32 | 0.004 | 3e−5 |

In almost all cases (only except for *email-Eu* at node level), for each dataset at each considered level, there is at least one positive ratio, indicating that there is more evidence of one heavy-tailed distribution than the exponential distribution.

Table 4. Loglikehood ratio when fitting the degree distribution to each of the following heavy-tailed distributions versus the exponential distribution. A positive value means there is more evidence of the heavy-tailed distribution, while a negative value means there is more evidence of exponential distribution. In almost all cases, for each dataset at each level, at least one heavy-tailed distribution has a positive ratio, indicating more evidence against the exponential distribution.

| Level | **Node** | | | **Edge** | | | **Triangle** | | | **4clique** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | pw | trunpw | lgnorm | pw | trunpw | lgnorm | pw | trunpw | lgnorm | pw | trunpw | lgnorm |
| coauth-DBLP | **1108** | **1108** | **1108** | **5616** | **5735** | **5718** | - | - | - | - | - | - |
| coauth-Geology | **10.77** | **11.3** | **11.3** | **122.1** | **123.3** | **123.4** | - | - | - | - | - | - |
| coauth-History | **429** | **430** | **429.9** | - | - | - | - | - | - | - | - | - |
| DAWN | -4.9 | -0.5 | **0.3** | **4025** | **4389** | **4303** | **1392** | **1426** | **1417** | **372.6** | **377.8** | **374.4** |
| email-Eu | -15.3 | -1.3 | -1.1 | **10.9** | **11.8** | **11.5** | **6.8** | **6.9** | **6.8** | -2 | **0.15** | -0.19 |
| NDC-classes | **2.17** | **18.9** | **14.3** | **44.9** | **44.9** | **44.9** | - | - | - | - | - | - |
| NDC-substances | -8 | **24.8** | **20.5** | **10.9** | **21.8** | **19.4** | **0.6** | **0.6** | **0.6** | - | - | - |
| tags-ask-ubuntu | -1.4 | **6.1** | **4.9** | **36.1** | **41.3** | **39.7** | **378.6** | **383.2** | **381** | **21.5** | **21.5** | **25.9** |
| tags-math | -11.4 | **0.37** | -1.1 | **20.4** | **24** | **23.6** | **96.4** | **100.8** | **99.3** | **107.5** | **107.5** | **112** |
| tags-stack-overflow | **202.8** | **245.1** | **241.8** | **394268** | **395917** | **395852** | **33198** | **33351** | **33319** | **31.6** | **31.6** | **31.6** |
| threads-ask-ubuntu | **2322** | **2330** | **2326** | - | - | - | - | - | - | - | - | - |
| threads-math | **67574** | **67751** | **67725** | **1524** | **1534** | **1528** | - | - | - | - | - | - |
| threads-stack-overflow | **2486** | **2549** | **2543** | **4760** | **4785** | **4775** | - | - | - | - | - | - |

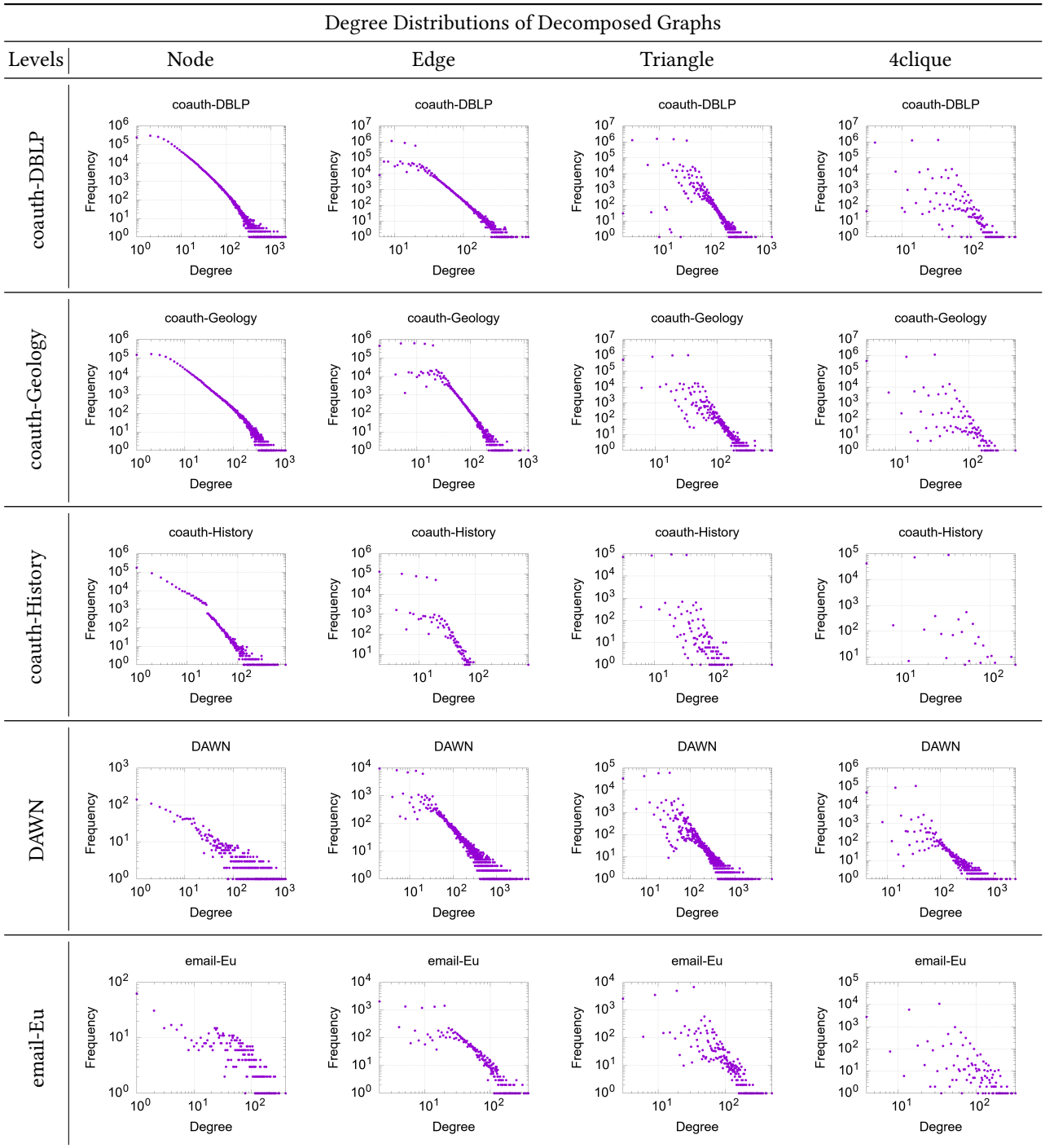


Fig. 3. Degree distributions of decomposed graphs at all decomposition levels.

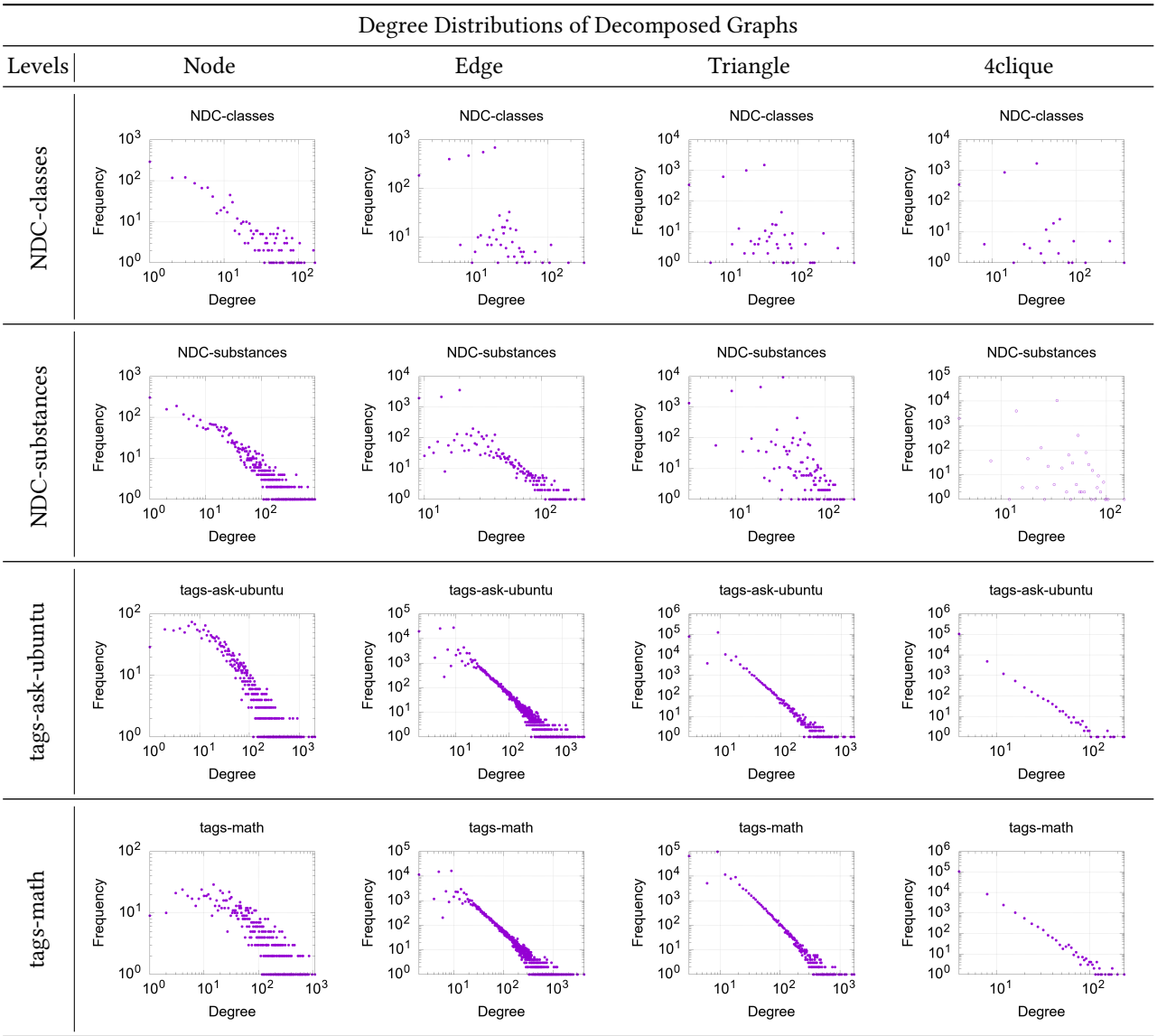


Fig. 4. Degree distributions of decomposed graphs at all decomposition levels (cont).

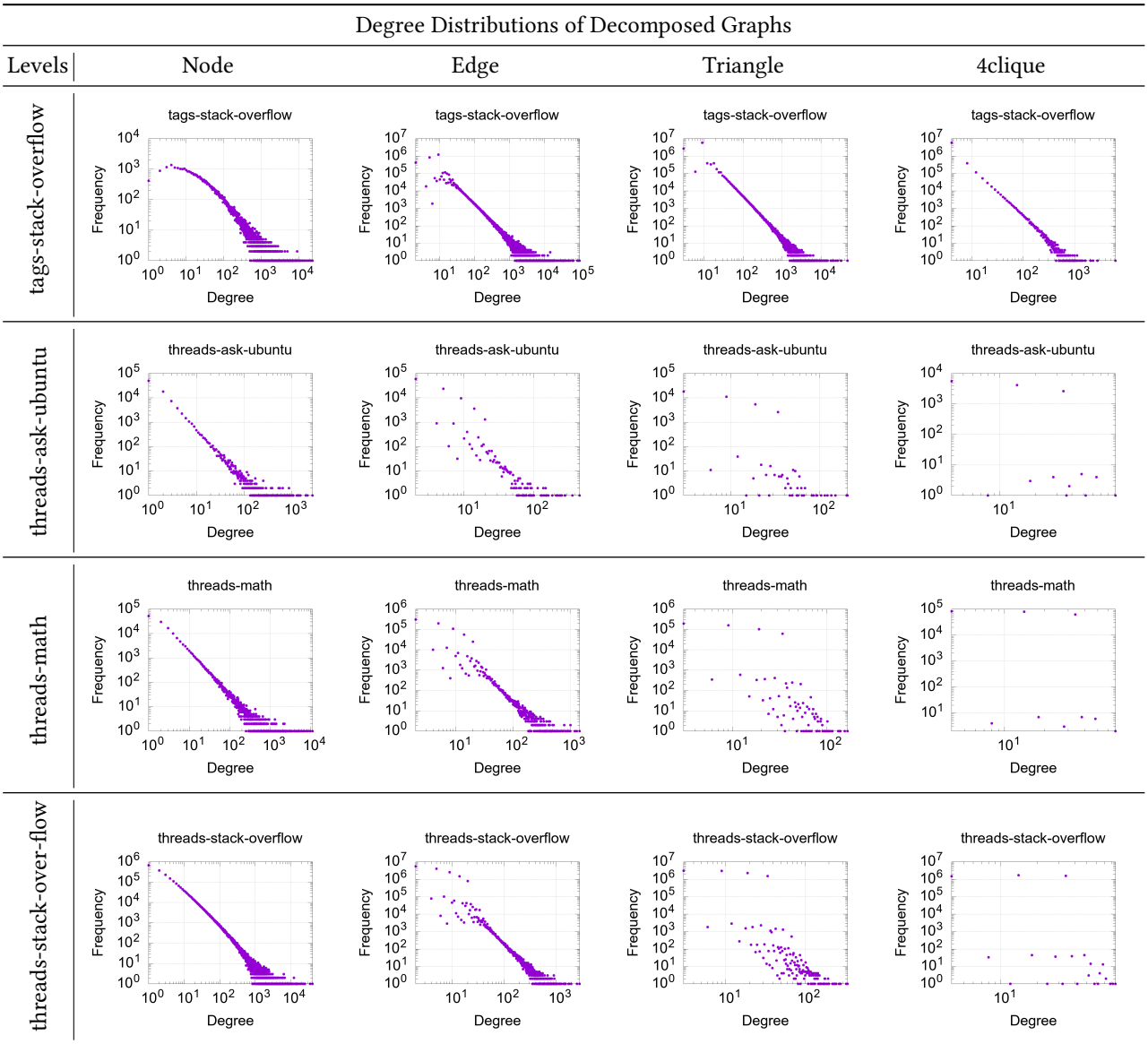


Fig. 5. Degree distributions of decomposed graphs at all decomposition levels (cont).

## B.3 Diameter and clustering coefficients

Effective diameters and global clustering coefficients are given in Tables 5 and 6.

Table 5. Properties of node-level decomposed graphs of all the datasets. The *connected component* column reports the proportion of total nodes involved in the largest connected component. The *diameter* and *clustering coefficient* are compared against a null model. Average and standard deviation of 10 random hypergraphs are reported. All node-level decomposed graphs possess a relatively small diameter. Almost all of them have clustering coefficients significantly higher than that of the null model.

| **Measure** | **# Nodes** | **Connected component** | **Diameter** | | **Clustering coefficient** | |
|---|---|---|---|---|---|---|
| Dataset | Real data | Real data | Real data | Null model | Real data | Null model |
| coauth-DBLP | 1,924,991 | 0.86 | 6.8 | 6.7 ±9e−3 | 0.60 | 0.31 ±1e−4 |
| coauth-Geology | 1,256,385 | 0.72 | 7.1 | 6.8 ±8e−3 | 0.57 | 0.42 ±2e−4 |
| coauth-History | 1,014,734 | 0.22 | 11.9 | 17 ±0.19 | 0.24 | 0.26 ±2e−4 |
| DAWN | 2,558 | 0.89 | 2.6 | 1.85 ±8e−5 | 0.64 | 0.30 ±9e−5 |
| email-Eu | 998 | 0.98 | 2.8 | 1.85 ±7e−5 | 0.49 | 0.36 ±5e−4 |
| NDC-classes | 1,161 | 0.54 | 4.6 | 2.6 ±6e−3 | 0.61 | 0.32 ±2e−3 |
| NDC-substances | 5,311 | 0.58 | 3.5 | 2.5 ±9e−3 | 0.40 | 0.17 ±6e−4 |
| tags-ask-ubuntu | 3,029 | 0.99 | 2.4 | 1.9 ±2e−5 | 0.61 | 0.14 ±7e−5 |
| tags-math | 1,629 | 0.99 | 2.1 | 1.8 ±1e−4 | 0.63 | 0.46 ±2e−4 |
| tags-stack-overflow | 49,998 | 0.99 | 2.7 | 1.9 ±2e−6 | 0.63 | 0.03 ±1e−6 |
| threads-ask-ubuntu | 125,602 | 0.65 | 4.7 | 11.9 ±0.042 | 0.11 | 0.19 ±7e−4 |
| threads-math | 176,445 | 0.86 | 3.7 | 4.9 ±4e−3 | 0.32 | 0.12 ±1e−4 |
| threads-stack-overflow | 2,675,995 | 0.86 | 4.5 | 5.9 ±2e−3 | 0.18 | 0.12 ±2e−5 |

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 |

## B.4 Singular value distribution

While the plots for distributions of singular values are provided at all 4 levels (Figures 6, 7, 8), note that the pattern of 'skewed singular-value distribution' is only inspected up to the decomposition level level in which the graph retains a 'giant connected component'. In addition, we also provide the loglikehood ratio of fitting empirical distributions to a heavy-tailed distribution versus the exponential distribution (Exp.) in Table 7:

- Power-law (pw) vs. Exp.
- Truncated power-law (trunpw) vs. Exp.
- Lognormal (lgnorm) vs. Exp.

In almost all cases (only except for *tags-stack-overflow* at edge level), for each dataset at each considered level, there is at least one positive ratio, indicating that there is more evidence of one heavy-tailed distribution than the exponential distribution. Due to underflow problems, in some cases the results for truncated power-law are not available.

Table 7. Loglikehood ratio when fitting the distribution of singular values to each of the following heavy-tailed distributions versus the exponential distribution. A positive value means there is more evidence of the heavy-tailed distribution, while a negative value means there is more evidence of exponential distribution. In almost all cases, for each dataset at each level, at least one heavy-tailed distribution has a positive ratio, indicating more evidence against the exponential distribution.

| Level | **Node** | | | **Edge** | | | **Triangle** | | | **4clique** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | pw | trunpw | lgnorm | pw | trunpw | lgnorm | pw | trunpw | lgnorm | pw | trunpw | lgnorm |
| coauth-DBLP | **3.4** | **3.6** | **6.4** | **1.3** | **1.3** | **4.5** | - | - | - | - | - | - |
| coauth-Geology | -2.3 | - | **11.3** | **122.1** | **123.3** | **123.4** | - | - | - | - | - | - |
| coauth-History | -1 | -0.07 | **0.3** | - | - | - | - | - | - | - | - | - |
| DAWN | **16.8** | **16.8** | **22** | **0.5** | **0.6** | **0.5** | **3.3** | **3.3** | **3.3** | **0.04** | **0.2** | **0.2** |
| email-Eu | -1.3 | -0.14 | **0.4** | **10.9** | **11.8** | **11.5** | -1.2 | -0.12 | **0.4** | -0.8 | -0.07 | **0.4** |
| NDC-classes | **1.2** | **1.3** | **1.3** | **1** | **1** | **3.8** | - | - | - | - | - | - |
| NDC-substances | **7.5** | **7.5** | **11.8** | **10.9** | - | **0.3** | -4 | -0.5 | **12.6** | - | - | - |
| tags-ask-ubuntu | **9.5** | **9.5** | **9.5** | -0.6 | **0.14** | **0.05** | -0.4 | **0.15** | **0.3** | -0.36 | -0.04 | **0.54** |
| tags-math | **9.9** | **9.9** | **9.9** | -1.3 | **0.01** | -0.1 | -0.03 | **0.001** | -0.001 | -0.06 | - | **0.13** |
| tags-stack-overflow | **6** | **6** | **6.1** | -1.5 | - | -0.15 | -0.5 | **0.1** | **0.1** | **31.6** | **31.6** | **31.6** |
| threads-ask-ubuntu | **2.3** | **2.3** | **5.7** | - | - | - | - | - | - | - | - | - |
| threads-math | **6.6** | **6.6** | **11.4** | **0.44** | **0.44** | **3** | - | - | - | - | - | - |
| threads-stack-overflow | **2.1** | **2.1** | **2.1** | -2.6 | -0.3 | **4.3** | - | - | - | - | - | - |

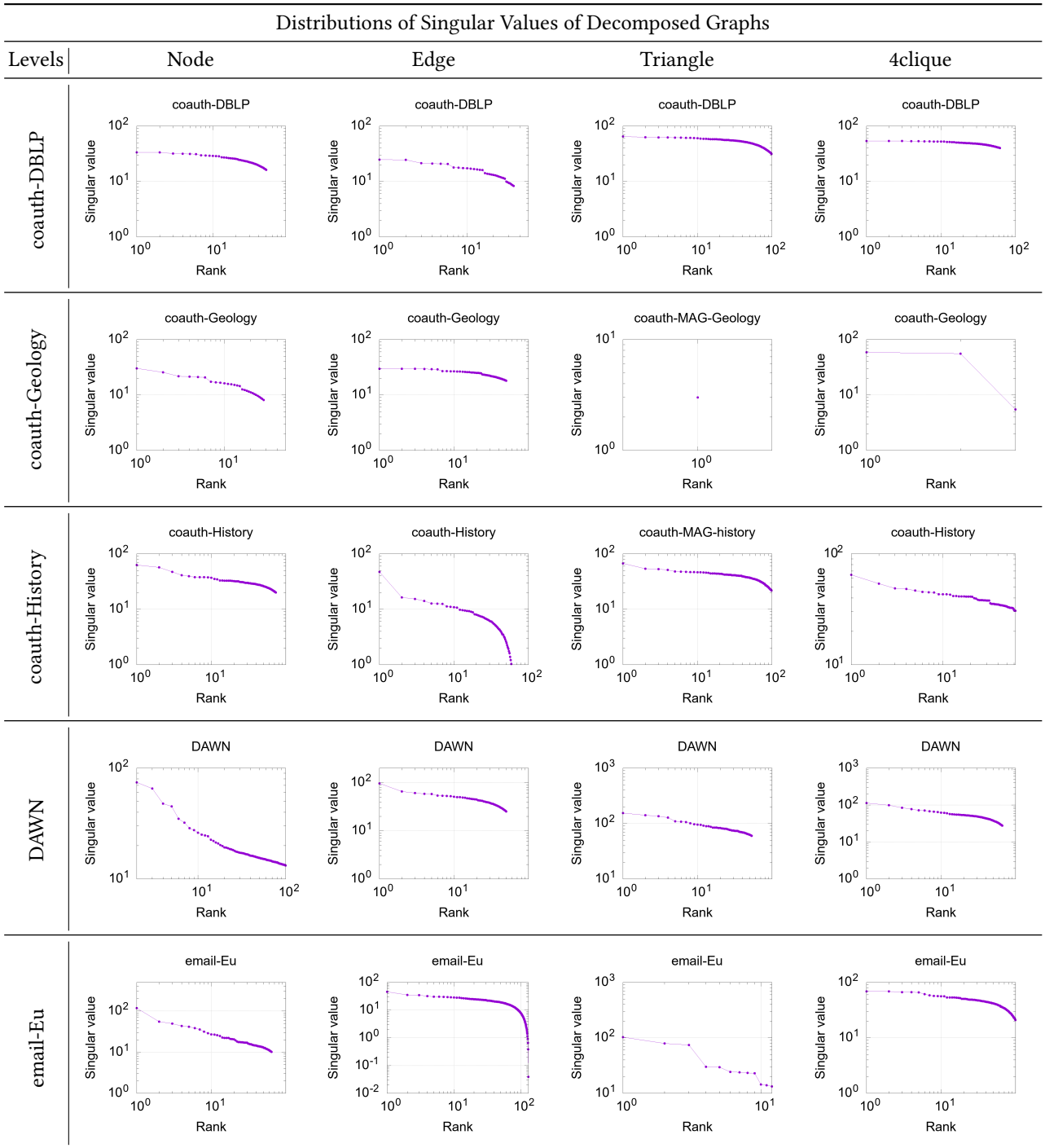


Fig. 6. Distributions of singular values of decomposed graphs at all decomposition levels.

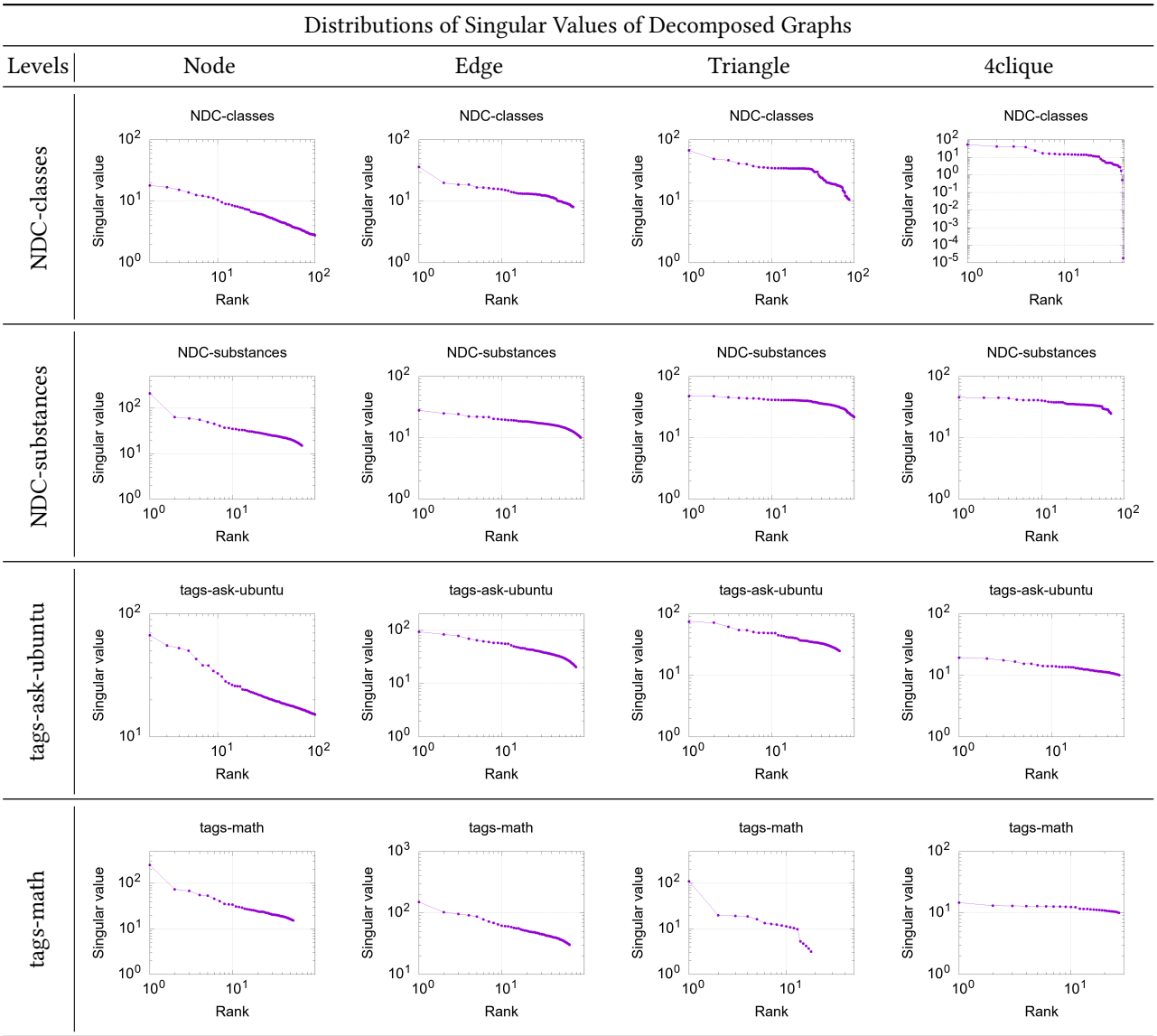


Fig. 7. Distributions of singular values of decomposed graphs at all decomposition levels (cont).

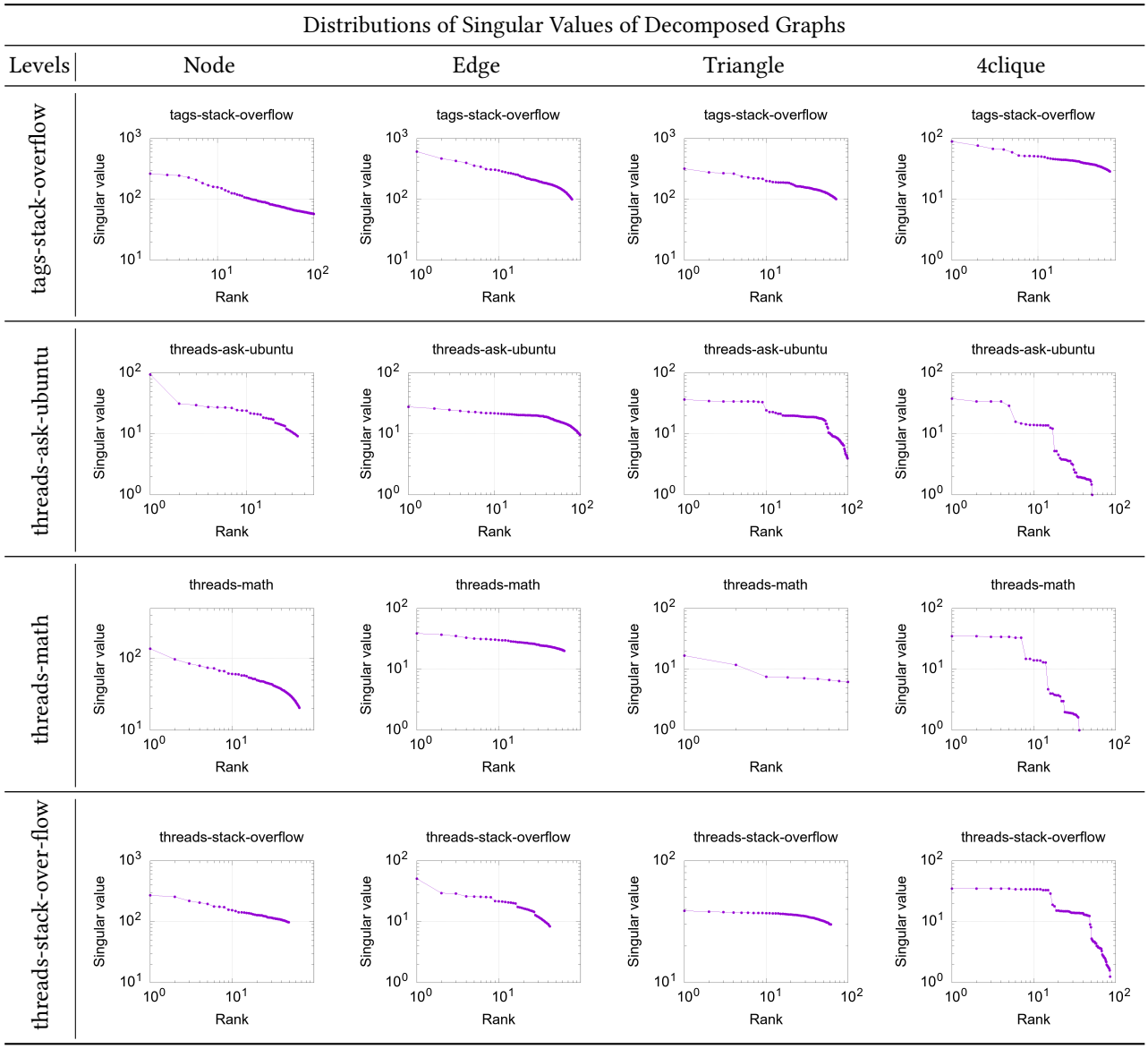


Fig. 8. Distributions of singular values of decomposed graphs at all decomposition levels (cont).

# 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 | **HyperPA (Proposed)** | Naive PA | Subset 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 |

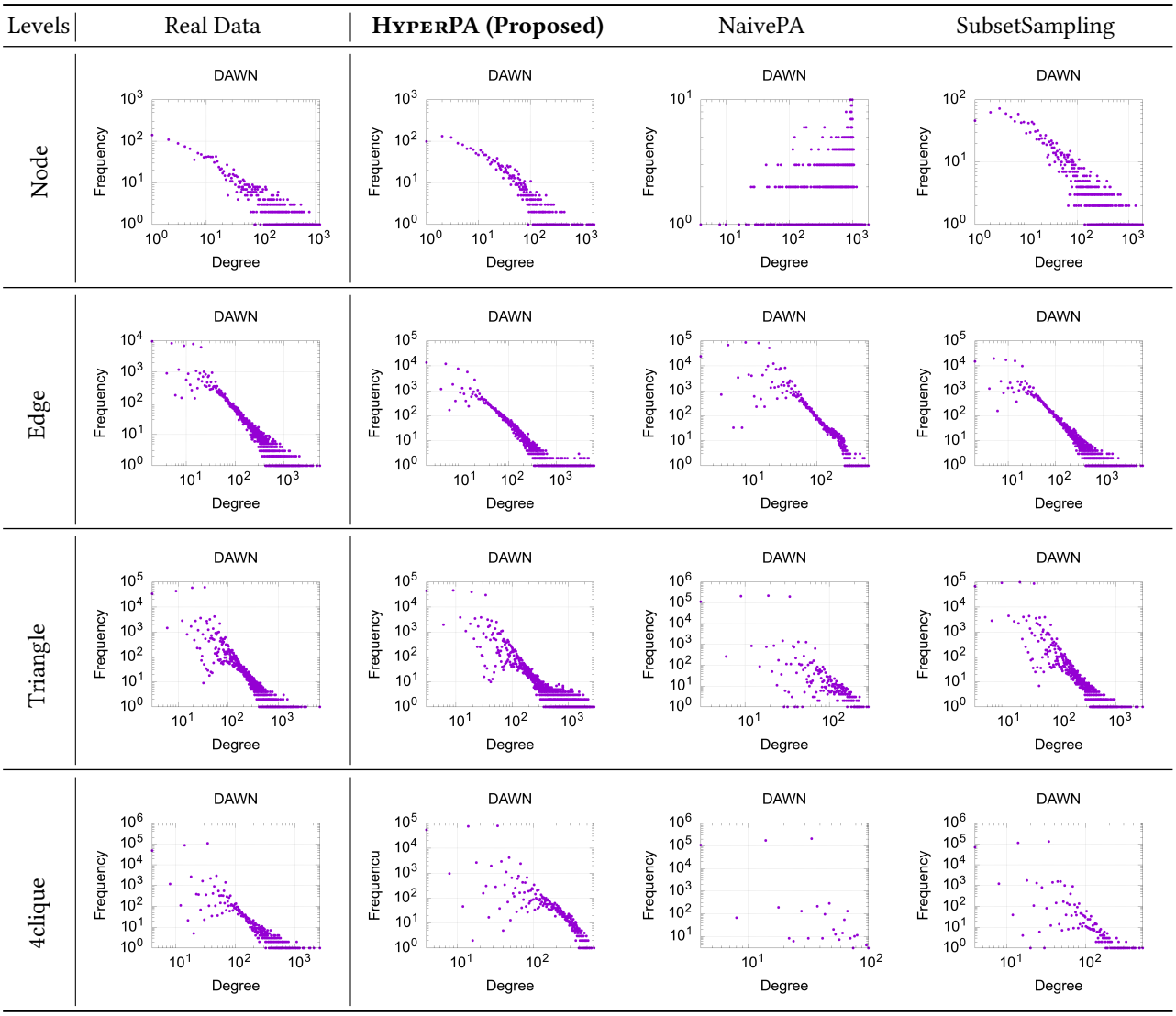


Fig. 9. Comparison of hypergraph generators with respect to degree distributions of decomposed graphs at different decomposition levels. The *DAWN* dataset was used to learn $S$, $NP$ and $n$.

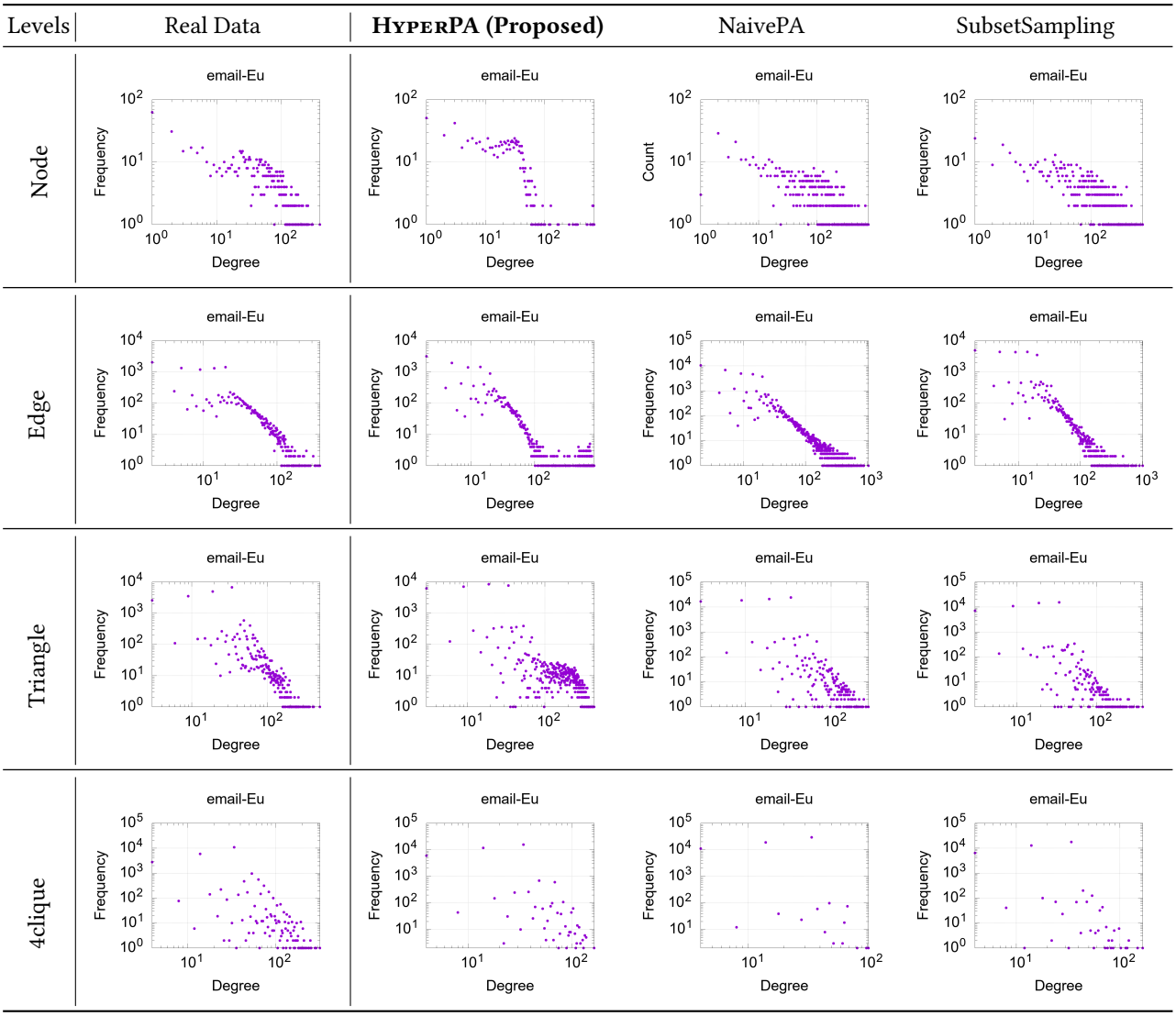


Fig. 10. Comparison of hypergraph generators with respect to degree distributions of decomposed graphs at different decomposition levels. The *email-Eu* dataset was used to learn $S$, $NP$ and $n$.

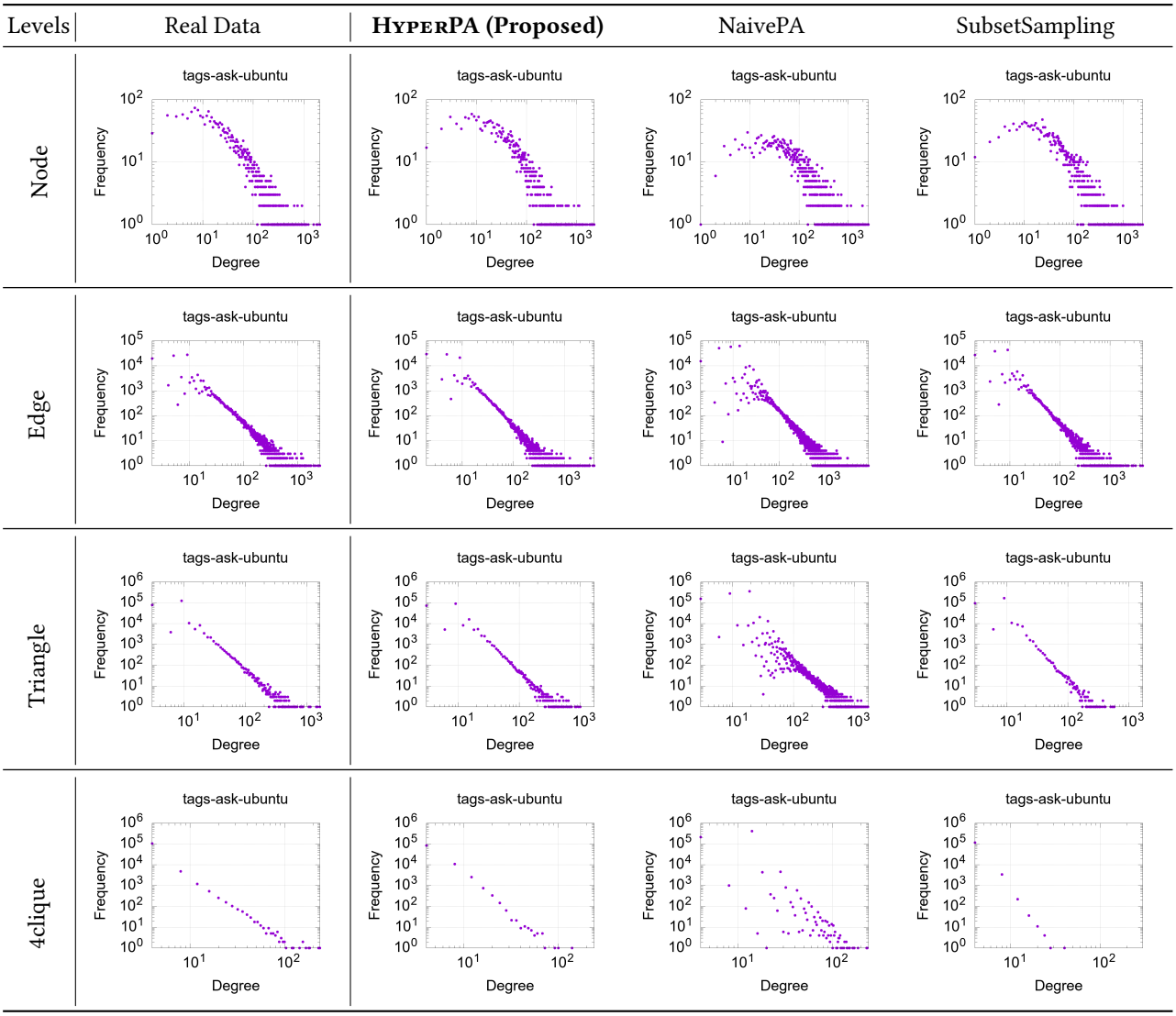


Fig. 11. Comparison of hypergraph generators with respect to degree distributions of decomposed graphs at different decomposition levels. The *tags-ask-ubuntu* dataset was used to learn $S$, $NP$ and $n$.

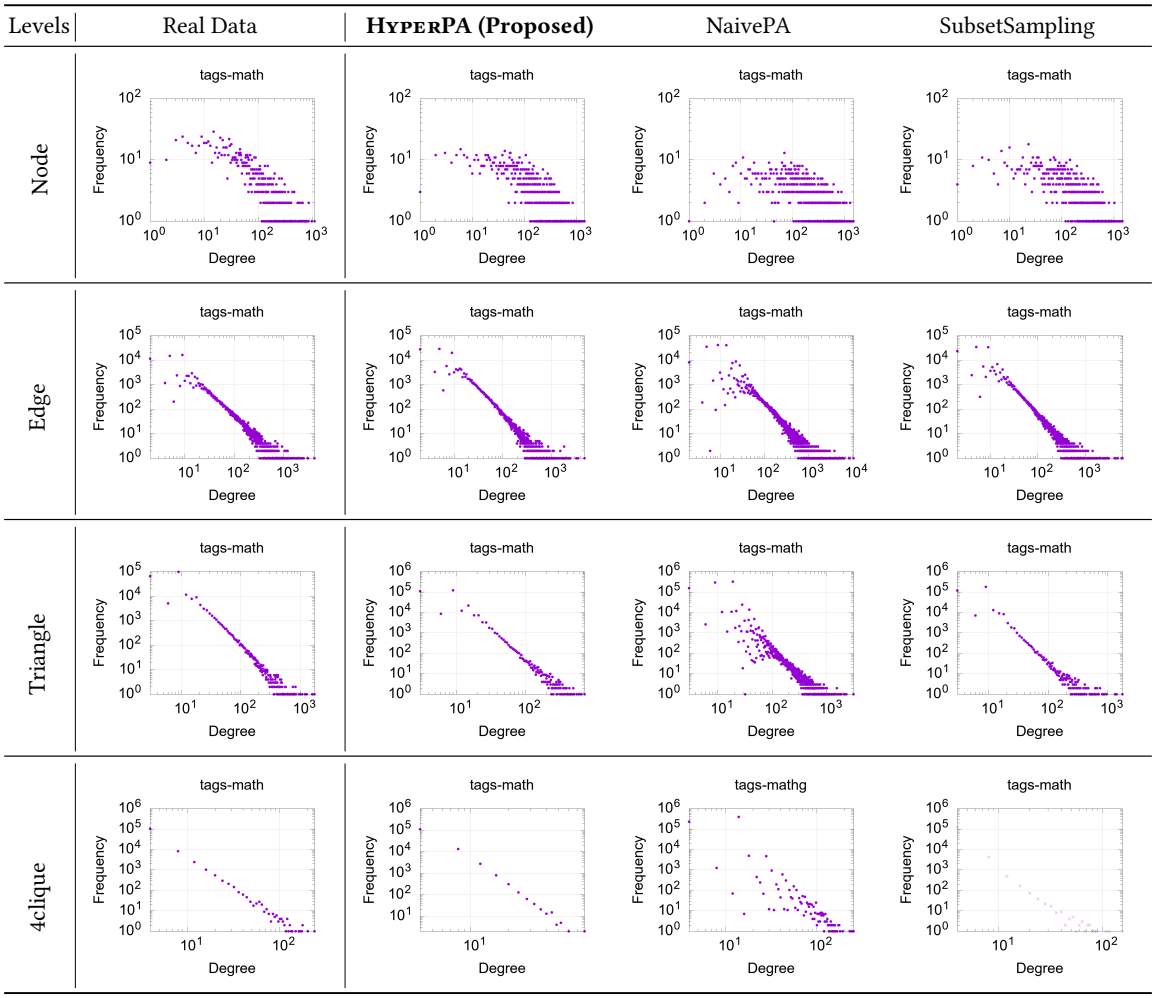


Fig. 12. Degree distributions of decomposed graphs at different decomposition levels of *tags-math*, both real and generated datasets. Those generated by HyperPA achieve more similartiy to the real distribution (confirmed in Table ??).

## 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 | **HyperPA (Proposed)** | Naive PA | Subset 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 |

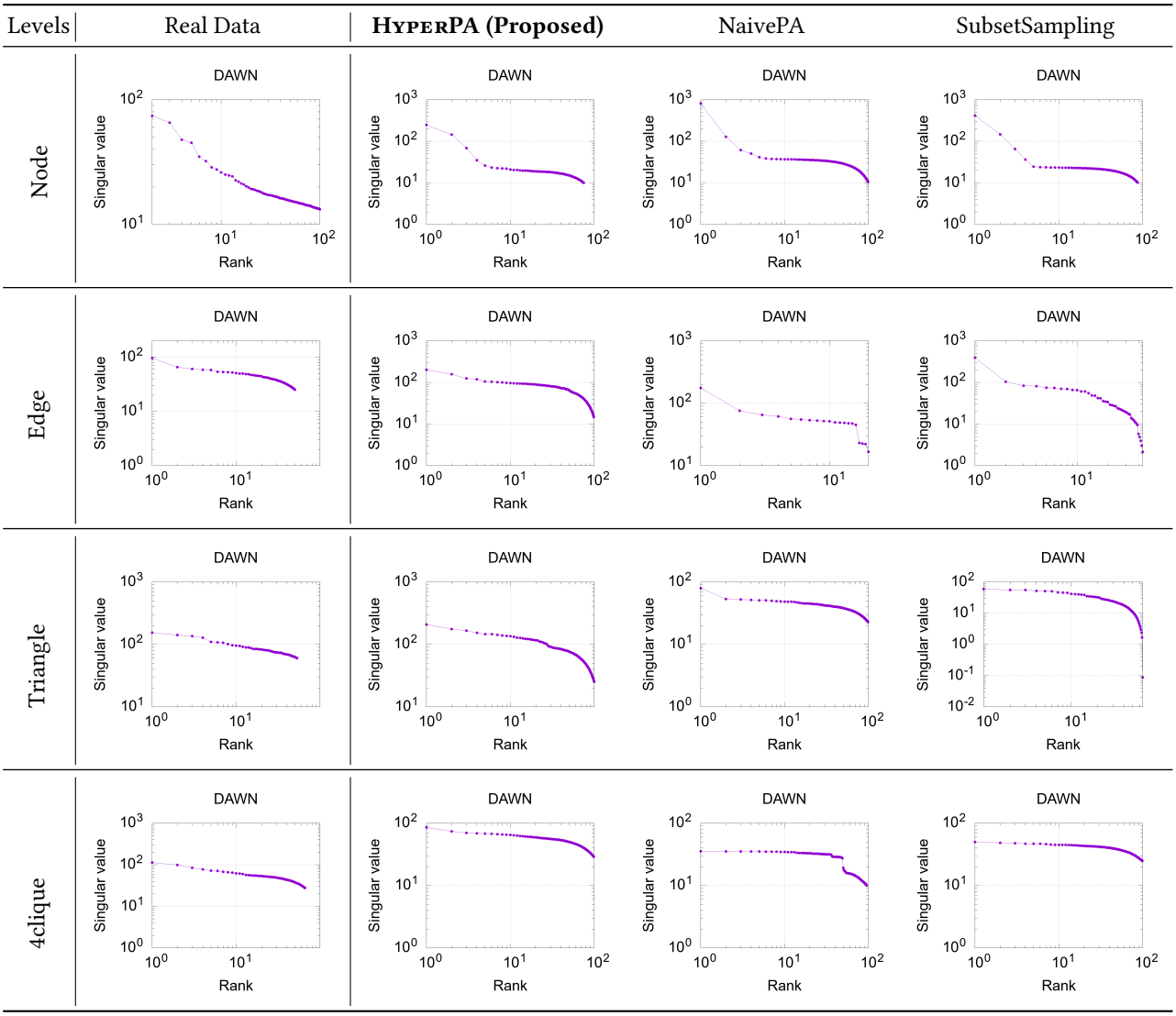


Fig. 13. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *DAWN* dataset was used to learn $S$, $NP$ and $n$.

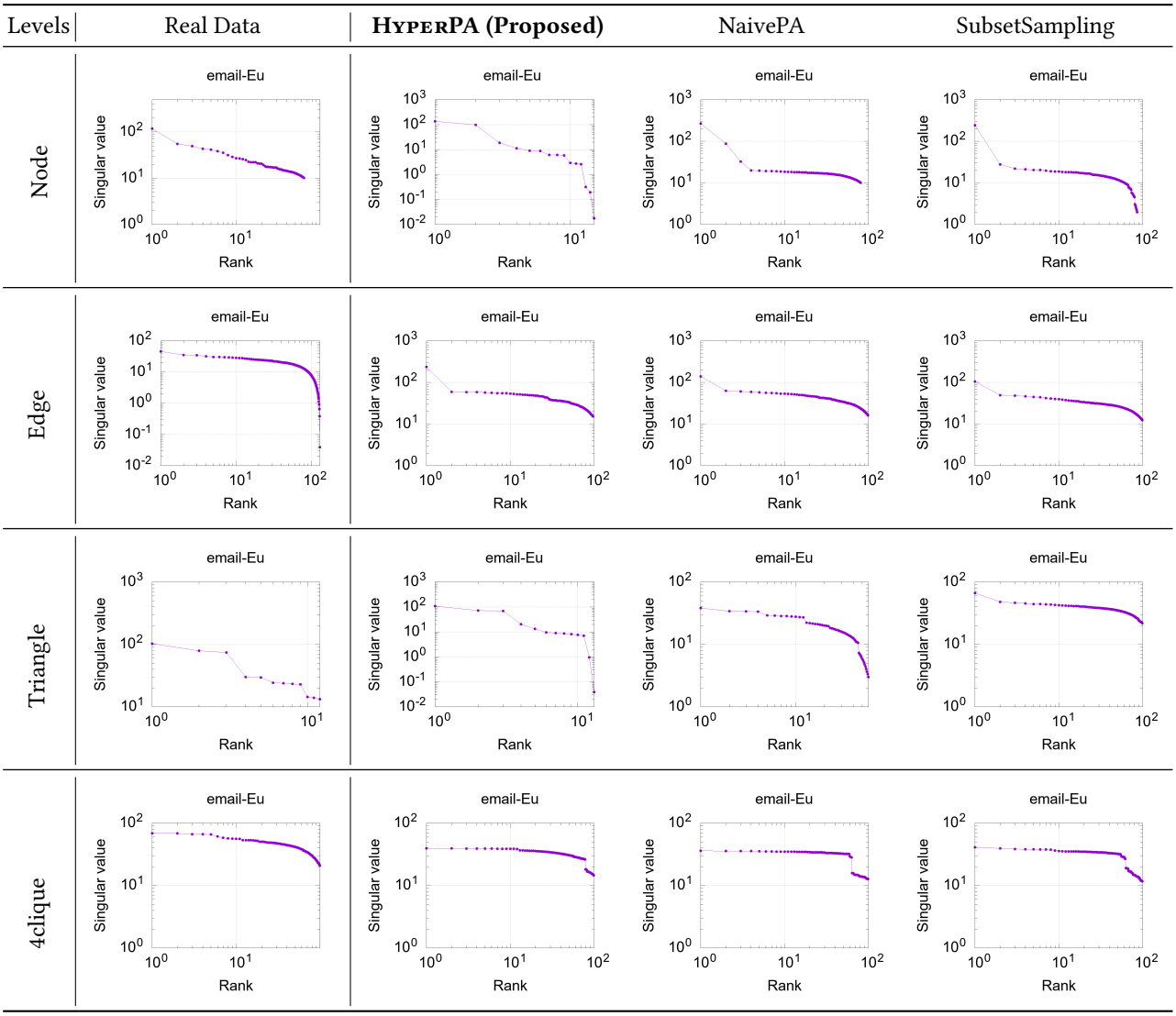


Fig. 14. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *email-Eu* dataset was used to learn $S$, $NP$ and $n$.

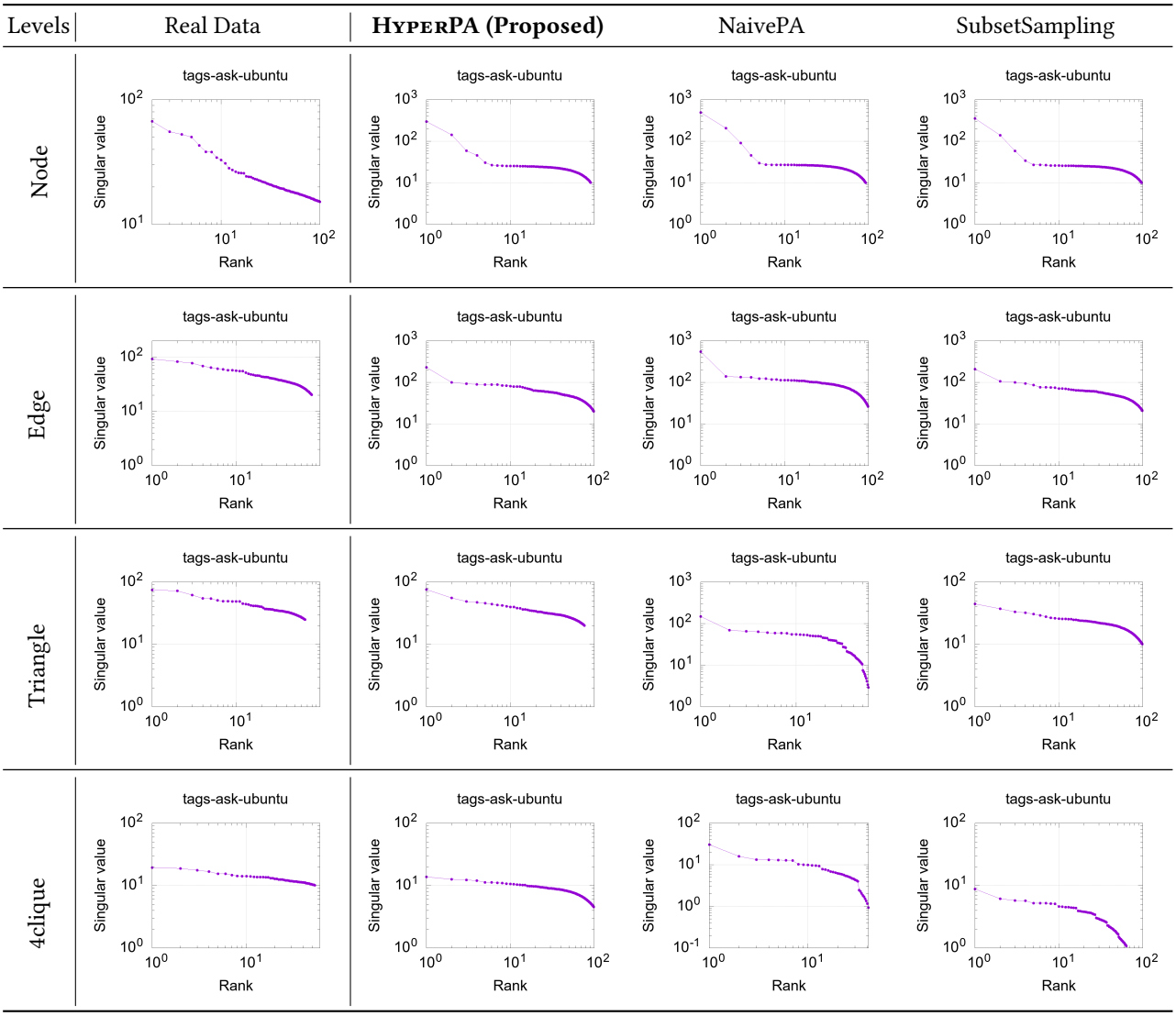


Fig. 15. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *tags-ask-ubuntu* dataset was used to learn $S$, $NP$ and $n$.

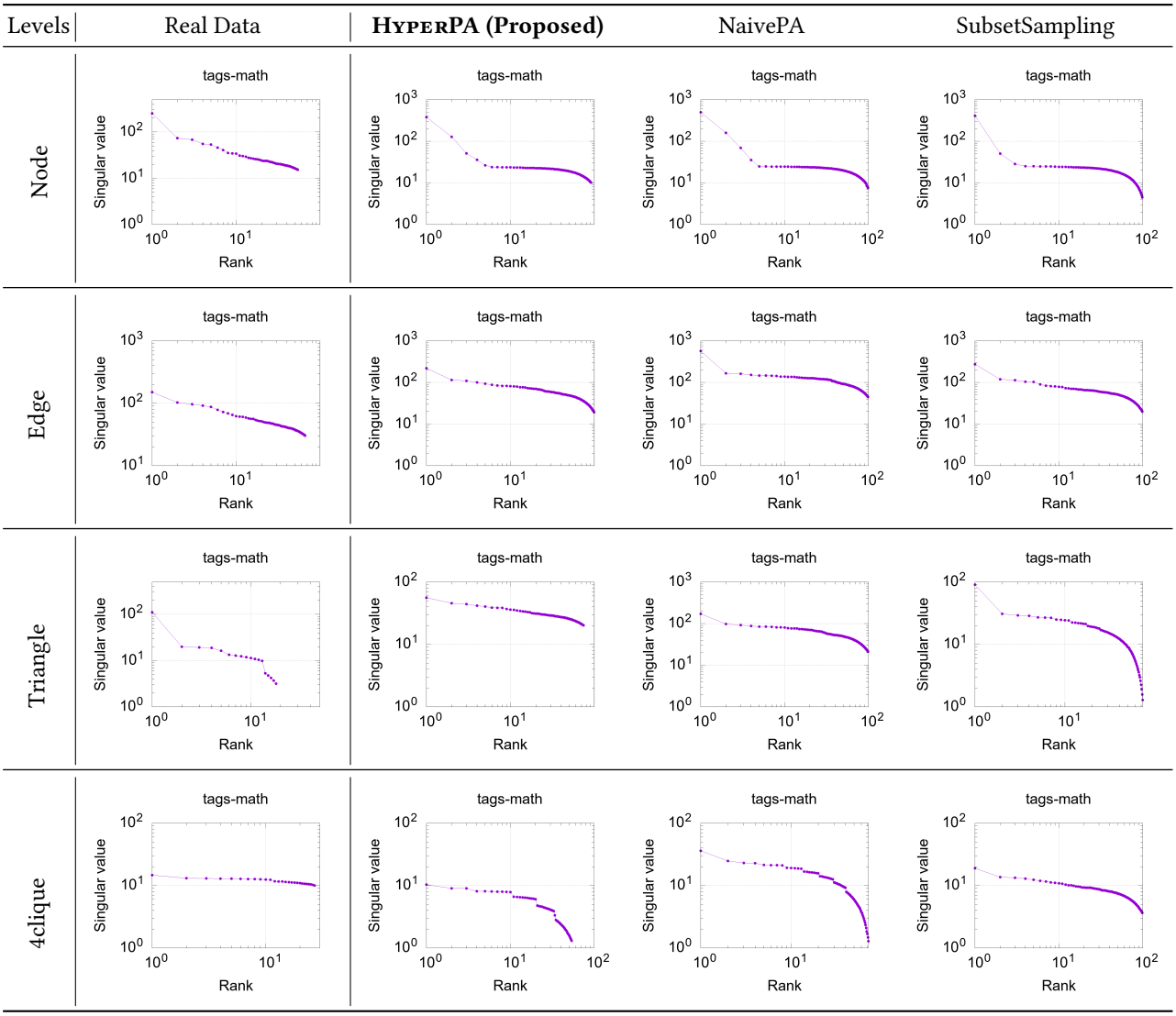


Fig. 16. Comparison of hypergraph generators with respect to singular-value distributions of decomposed graphs at different decomposition levels. The *tags-math* dataset was used to learn $S$, $NP$ and $n$.