Supplementary Information

Clustering Coefficient Reflecting Pairwise Relationships within Hyperedges

Rikuya Miyashita[1]**, Shiori Hironaka**[2,*]**, and Kazuyuki Shudo**[2]

[1]Tokyo Institute of Technology, Department of Mathematical and Computing Science, Tokyo, 152-8552, Japan
[2]Kyoto University, Academic Center for Computing and Media Studies, Kyoto, 606-8501, Japan
[*]hironaka@media.kyoto-u.ac.jp

Extended Analysis of Clustering Coefficients in Hypergraphs

This supplementary document extends the analysis of the proposed clustering coefficient presented in the main paper. While our primary paper focused on three datasets (email-Enron, primary-school, and NDC-classes[1]), here we provide a more comprehensive evaluation by including five additional real-world hypergraph datasets: DavisClub, congress-bills, threads-ask-ubuntu, threads-math-sx, and DAWN[2]. These additional datasets represent diverse domains: DavisClub captures the classic Southern Women social participation study with interactions between women and events; congress-bills documents co-sponsorship patterns among US legislators; threads-ask-ubuntu and threads-math-sx represent user interactions in online Q&A platforms; and DAWN records co-occurring drug usage reported in emergency room visits. All these additional datasets are publicly available[2].

Table S1 summarizes the key statistics of all eight datasets analyzed in this study, including the original three from the main paper and the five additional datasets. Figure S1 illustrates the distribution of order-3 motifs across all datasets, providing insight into their structural composition. The datasets vary significantly in size and density: from the small DavisClub network (18 nodes, 14 hyperedges) to the much larger threads-math-sx (176,445 nodes, 719,792 hyperedges). They also exhibit different average hyperedge sizes ($\bar{s}$), with DAWN having relatively small hyperedges (average size 1.58) and NDC-classes featuring larger ones (average size 6.97).

We present a comprehensive analysis of the proposed clustering coefficient on all eight datasets. The average clustering coefficients are shown in Table S1. Figures S2 and S3 show the distribution of clustering coefficients, and Table S2 shows the correlation coefficients. Table S3 shows the number of nodes assigned clustering coefficients of either 0 or 1. Due to their higher computational time complexity compared to the proposed definition, existing definitions could not be computed on large datasets; therefore, the parts that could not be calculated are left blank.

Our extended analysis reveals that the findings from the main paper are largely consistent across this diverse range of hypergraph structures, with one notable exception: the DavisClub dataset. For most datasets, our proposed method shows moderate to strong correlation with existing methods (correlation coefficients ranging from 0.599 to 0.998, as shown in Table S2), indicating that it captures similar clustering tendencies. However, for the DavisClub dataset, our method produces substantially different results compared to existing definitions ($\rho_{ZP} = -0.513$, indicating negative correlation). This divergence can be explained by the structural characteristics of DavisClub, which contains predominantly large hyperedges. Our proposed method assigns weaker relationship weights to node pairs that belong to large hyperedges, resulting in lower clustering coefficient values even when these nodes form triangular structures. In contrast, existing methods treat all triangles as indicators of clustering regardless of hyperedge size, leading to higher coefficient values. This distinction highlights a key advantage of our approach: the ability to differentiate between strong clustering (represented by small, focused hyperedges) and weaker clustering (represented by large, more diffuse hyperedges).

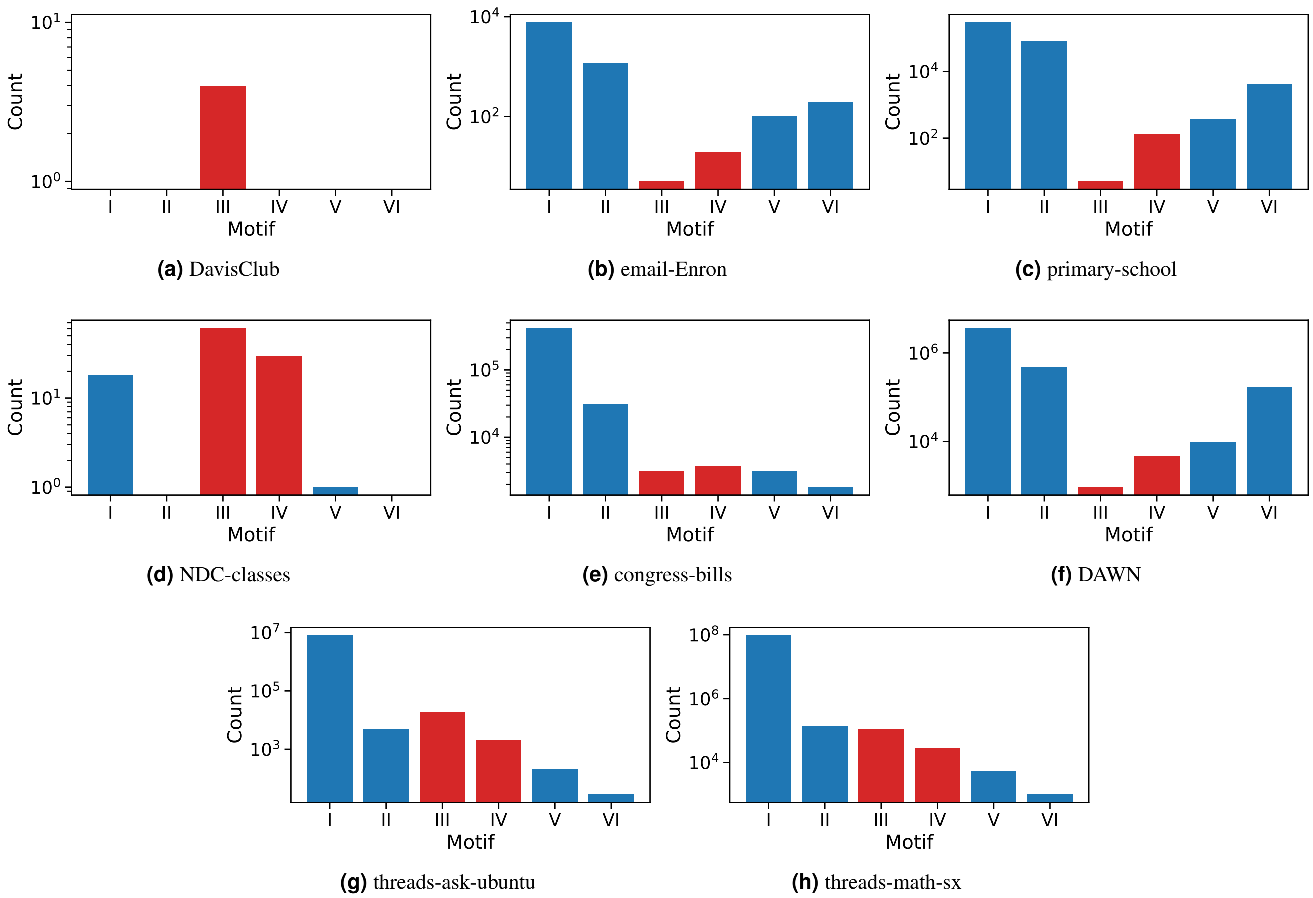
Impact of Hyperedge Size on Clustering Coefficients

We analyze the relationship between hyperedge size and the value of the clustering coefficients to understand how different definitions respond to variations in hyperedge structure. For each node, we calculated the average size of hyperedges containing that node (denoted as mean_s) and examined how clustering coefficient values vary with this metric across all datasets.

Figure S4 illustrates the relationship between mean hyperedge size and clustering coefficients for all four definitions. The results reveal two key findings. First, our proposed clustering coefficient tends to decrease as average hyperedge size increases. This reflects the core principle of our approach: pairwise relationships within larger hyperedges should receive lower weights since the connection strength between any two nodes diminishes as the number of nodes in the hyperedge increases. This

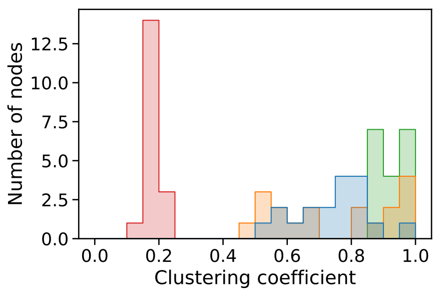
Table S1. Dataset statistics. N: number of nodes, M: number of hyperedges, $\mathscr{M}$: number of edges in the corresponding bipartite graph, $\bar{k}$: average degree of the node, $\bar{s}$: average size of the hyperedge, and $\bar{C}_{\text{Opsahl}}$, $\bar{C}_{\text{Zhou}}$, $\bar{C}_{\text{proposed}}$, and $\bar{C}_{\text{baseline}}$ are the averages of $C_{\text{Opsahl}}(v)$, $C_{\text{Zhou}}(v)$, $C_{\text{proposed}}(v)$, and $C_{\text{baseline}}(v)$ for all nodes v.

Dataset	N	M	$\mathscr{M}$	$\bar{k}$	$\bar{s}$	$\bar{C}_{\text{Opsahl}}$	$\bar{C}_{\text{Zhou}}$	$\bar{C}_{\text{proposed}}$	$\bar{C}_{\text{baseline}}$
DavisClub[2]	18	14	89	4.94	6.36	0.747	0.754	0.179	0.937
email-Enron[1,3]	143	1512	4550	31.82	3.01	0.684	0.527	0.411	0.593
primary-school[1,4]	242	12 704	30 729	126.98	2.42	0.699	0.674	0.517	0.526
NDC-classes[1]	628	816	5688	9.06	6.97	0.313	0.141	0.226	0.772
congress-bills[2]	1718	260 851	954 262	555.45	3.66	-	0.708	0.184	0.755
DAWN[2]	2558	2 272 433	3 596 794	1406.10	1.58	-	0.190	0.481	0.638
threads-ask-ubuntu[2]	125 602	192 947	346 537	2.76	1.80	0.021	0.018	0.080	0.175
threads-math-sx[2]	176 445	719 792	1 610 393	9.13	2.24	0.134	0.125	0.175	0.368

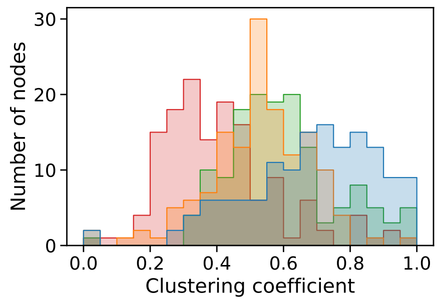


(a) DavisClub **(b)** email-Enron **(c)** primary-school **(d)** NDC-classes **(e)** congress-bills **(f)** DAWN **(g)** threads-ask-ubuntu **(h)** threads-math-sx

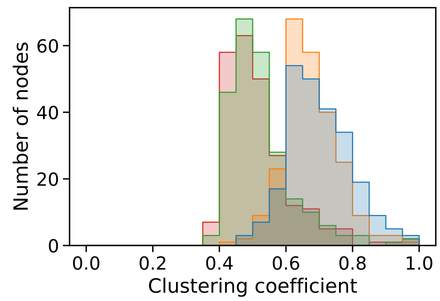
Figure S1. Number of higher-order motifs of order 3 for each dataset.



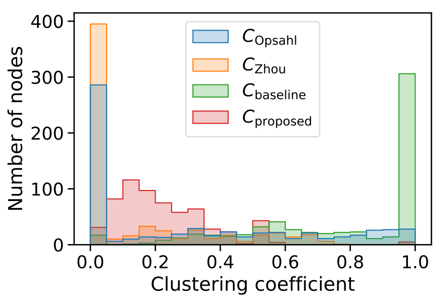
(a) DavisClub



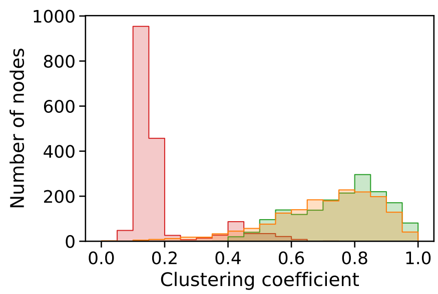
(b) email-Enron



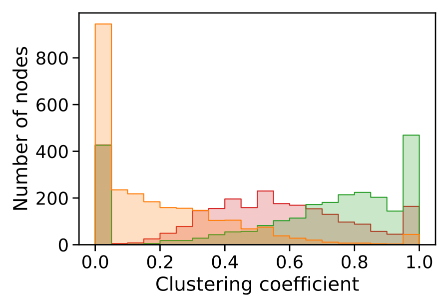
(c) primary-school



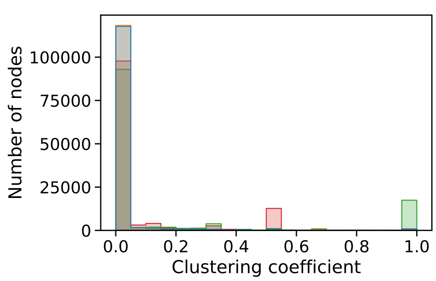
(d) NDC-classes



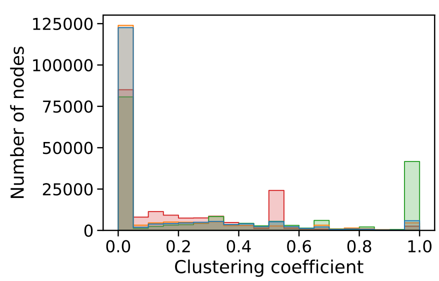
(e) congress-bills



(f) DAWN



(g) threads-ask-ubuntu

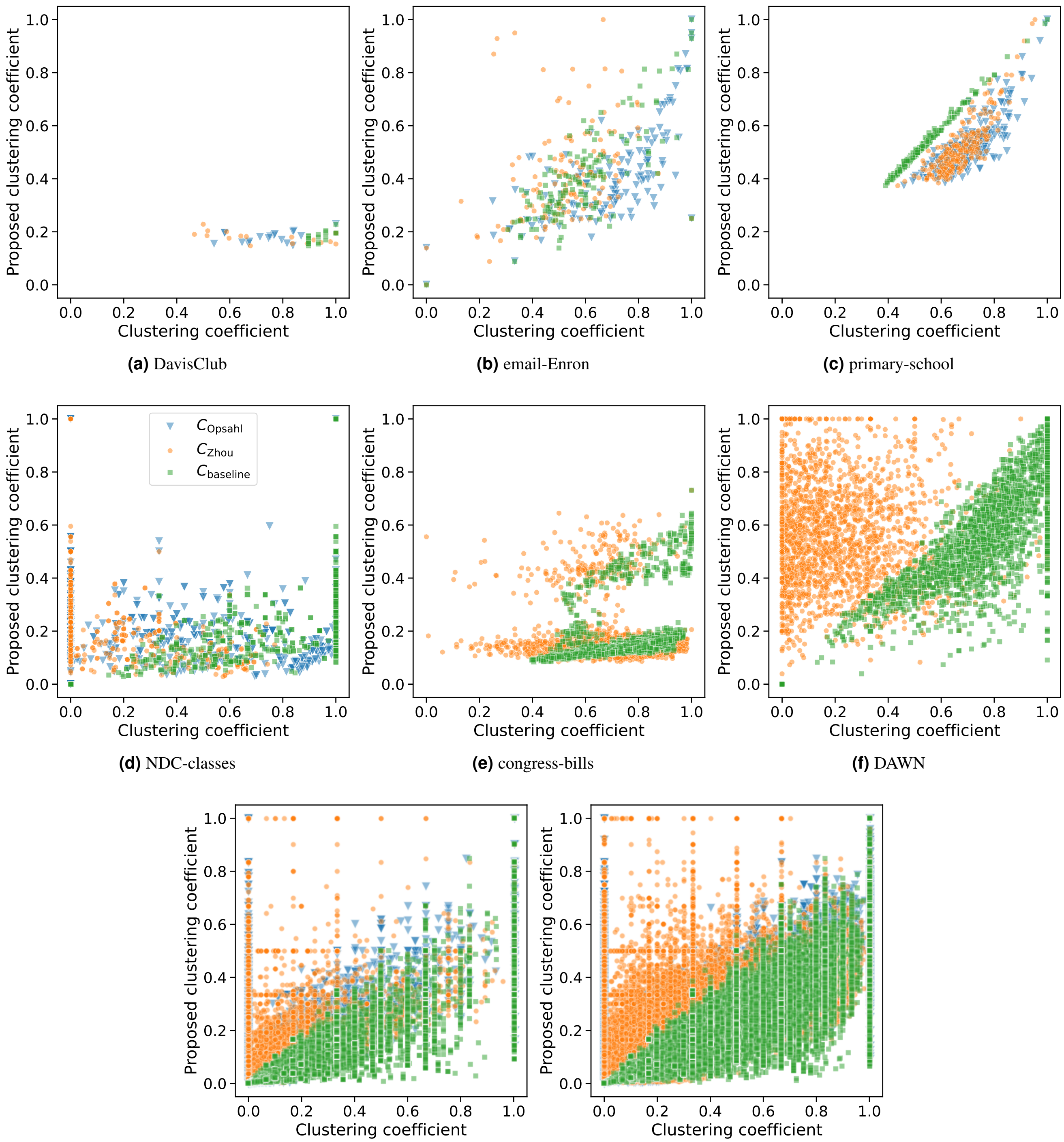


(h) threads-math-sx

Figure S2. The histogram of the clustering coefficients.

Table S2. Correlation coefficients of the clustering coefficients.

Dataset	ρ_{OP}	ρ_{ZP}	ρ_{SP}
DavisClub	0.411	-0.513	0.700
email-Enron	0.702	0.346	0.783
primary-school	0.827	0.840	0.998
NDC-classes	-0.359	-0.408	0.599
congress-bills		-0.141	0.372
DAWN		0.281	0.897
threads-ask-ubuntu	0.285	0.257	0.953
threads-math-sx	0.377	0.330	0.917



(a) DavisClub **(b)** email-Enron **(c)** primary-school

(d) NDC-classes **(e)** congress-bills **(f)** DAWN

(g) threads-ask-ubuntu **(h)** threads-math-sx

Figure S3. The scatter plot of the clustering coefficients.

Table S3. Number of nodes with clustering coefficient of 0 or 1 across different datasets.

Dataset	Clustering Coefficient = 0				Clustering Coefficient = 1			
	C_{Opsahl}	C_{Zhou}	C_{proposed}	C_{baseline}	C_{Opsahl}	C_{Zhou}	C_{proposed}	C_{baseline}
DavisClub	0	0	0	0	1	3	0	3
email-Enron	2	2	1	1	4	1	1	4
primary-school	0	0	0	0	1	0	1	1
NDC-classes	283	385	17	17	9	1	5	303
congress-bills	-	1	0	0	-	3	0	4
DAWN	-	637	426	426	-	39	144	361
threads-ask-ubuntu	115606	115723	91749	91749	908	603	597	17426
threads-math-sx	122074	122649	80413	80413	5837	4539	2551	41599

behavior is consistently observed across all datasets. Second, existing definitions do not necessarily show decreasing values with increasing hyperedge size. The baseline definition, in particular, tends to maintain high coefficient values regardless of hyperedge size, as it treats all nodes within a hyperedge as strongly connected without considering the diluting effect of hyperedge size. Similarly, Opsahl's and Zhou's definitions focus on hyperedge overlap patterns rather than internal structure, making them less sensitive to changes in hyperedge size.

The inverse relationship between hyperedge size and clustering coefficient values in our proposed method aligns with intuitive expectations in many real-world networks. For instance, in social networks, being part of a small, focused group typically indicates stronger connections than membership in a large, diffuse community. Our definition effectively captures this intuition by properly weighting relationships based on the size of their containing hyperedges. This analysis demonstrates that considering hyperedge size is crucial for accurately measuring local density in complex networks, particularly those characterized by heterogeneous hyperedge sizes. Our proposed method offers a more defailed perspective that better reflects the varying strength of relationships within hypergraphs.

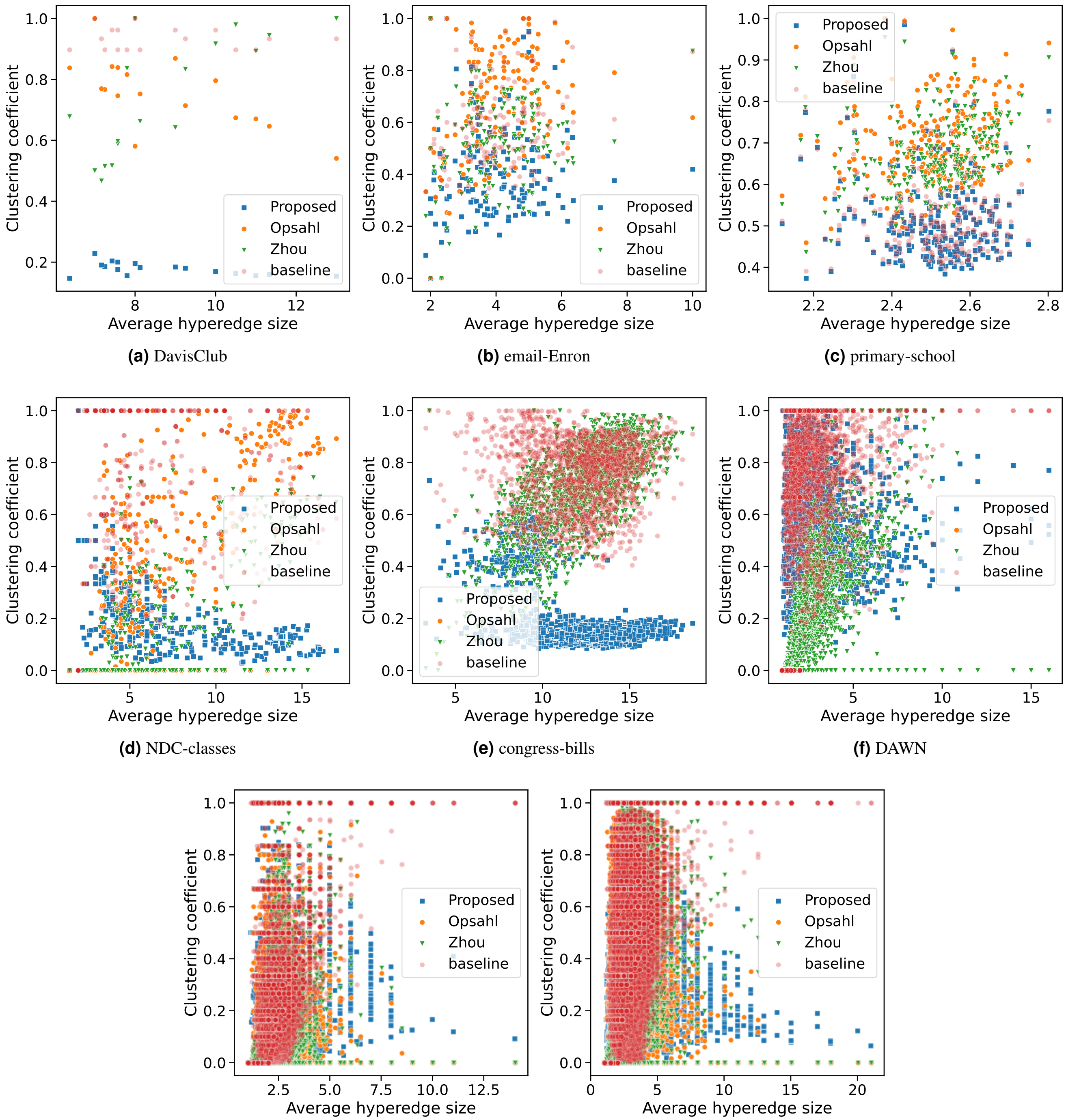
References

1. Benson, A. R., Abebe, R., Schaub, M. T., Jadbabaie, A. & Kleinberg, J. Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci.* **115**, E11221–E11230 (2018).

2. Lotito, Q. F., Musciotto, F., Montresor, A. & Battiston, F. Higher-order motif analysis in hypergraphs. *Commun. Phys.* **5**, 79 (2022).

3. Klimt, B. & Yang, Y. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, 217–226 (2004).

4. Stehlé, J. *et al.* High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one* **6**, e23176 (2011).



(a) DavisClub **(b)** email-Enron **(c)** primary-school

(d) NDC-classes **(e)** congress-bills **(f)** DAWN

(g) threads-ask-ubuntu **(h)** threads-math-sx

Figure S4. The relationship between the clustering coefficient and average hyperedge size containing the node.