Random Hypergraph Model Preserving Two-mode Clustering Coefficient

Rikuya Miyashita, Kazuki Nakajima, Mei Fukuda, and Kazuyuki Shudo

Tokyo Institute of Technology

Abstract. Real-world complex systems often involve interactions among more than two nodes, and such complex systems can be represented by hypergraphs. Comparison between a given hypergraph and randomized hypergraphs that preserve specific properties reveal effects or dependencies of the properties on the structure and dynamics. In this study, we extend an existing family of reference models for hypergraphs to generate randomized hypergraphs that preserve the pairwise joint degree distribution and the degree-dependent two-mode clustering coefficient of the original hypergraph. Using empirical hypergraph data sets, we numerically show that the extended model preserves the properties of the node and hyperedge as designed.

Keywords: Hypergraph · Two-mode clustering coefficient · Configuration model

1 Introduction

Networks are often used to represent complex systems that consist of nodes and pairwise interactions (i.e., edges) among the nodes [4]. On the other hand, real-world complex systems often involve interactions among more than two nodes [2]. For example, in email networks, there are multiple senders and receivers of a single e-mail [7]; in co-authorship networks, there are more than two coauthors for a single paper [11, 15]. Such complex systems can be represented as hypergraphs consisting of a set of nodes and hyperedges, where each hyperedge contains an arbitrary number of nodes.

Randomized networks that preserve specified properties are often used for analyzing the structure and dynamics of empirical networks [12, 8, 14]. In general, by comparing the structure or dynamics between a given network and randomized networks that preserve specified properties of the original network, we investigate how the preserved properties affect the structure or dynamics of interest of the original network [5]. The dK-series [6, 8, 14] is a family of models to generate such randomized networks. Given a network, the dK-series generates randomized networks that preserve up to the degree of each node, the pairwise joint degree distribution, and the degree-dependent clustering coefficient. A recent study proposed the hyper dK-series, which is an extension of the dK-series to the case of hypergraphs [10]. The hyper dK-series preserves up to the degree distribution of the node, the pairwise joint degree distribution of the node, the

Proc. 25th Int'l Conf. on Big Data Analytics and Knowledge Discovery (DaWaK 2023), August 2023

degree-dependent redundancy coefficient of the node, and the size distribution of the hyperedge in the original hypergraph.

In this study, we extend the hyper dK-series to preserve the pairwise joint degree distribution of the node and the degree-dependent two-mode clustering coefficient of the node, proposed in Ref. [13]. Using empirical hypergraph data sets, we numerically show that the extended hyper dK-series preserves the properties of a given hypergraph as designed. This paper is an extended version of our previous work published as an extended abstract [9]. This paper presents definitions and notations in Section 2 and further analysis and consideration of the experimental results in Section 4 and Fig. 1.

2 Preliminaries

We represent an unweighted hypergraph that consists of a set of nodes $V = \{v_1, \ldots, v_N\}$ and a set of hyperedges $E = \{e_1, \ldots, e_M\}$, where N is the number of nodes and M is the number of hyperedges. Then, we denote by $G = (V, E, \mathcal{E})$ the bipartite graph that corresponds to the given hypergraph, where $\mathcal{E}$ is a set of edges in the bipartite graph. An edge (v_i, e_j) is connected to each node v_i and each hyperedge e_j if and only if v_i belongs to the hyperedge e_j in the hypergraph. We assume that G does not contain multiple edges.

We denote by k_i and s_j the degree of node v_i (i.e., the number of hyperedges to which v_i belongs) and the size of hyperedge e_j (the number of nodes that belong to the hyperedge e_j), respectively. We also denote by $\bar{k}$ and $\bar{s}$ the average degree of the node and the average size of the hyperedge, respectively.

We use the joint degree distribution and the average degree of the nearest neighbors of nodes with degree k, which were defined in Ref. [10]. We denote by $P(k, k')$ the joint degree distribution of the node [10]. We denote by $k_{\mathrm{nn}}(k)$ by the average degree of the nearest neighbors of nodes with degree k [10].

We define the two-mode clustering coefficient of each node v_i [13]:

$$c_i = \frac{(\text{number of closed 4-paths centered on node } v_i)}{(\text{number 4-paths centered on node } v_i)}.$$

We denote by $\bar{c}$ the average two-mode clustering coefficient of the node. We define $c(k)$ the degree-dependent two-mode clustering coefficient of the node as

$$c(k) = \frac{1}{N(k)} \sum_{i=1, k_i=k}^{N} c_i,$$

where $N(k)$ is the number of nodes with degree k.

3 Extending the hyper dK-series to the case of $d_v = 2.5+$

The hyper dK-series generates a bipartite graph that preserves the joint degree distributions of the node in the subgraphs of size $d_v \in \{0, 1, 2, 2.5\}$ or less and

Table 1. Data sets. N: number of nodes, M: number of hyperedges, $\mathcal{M}$: number of edges in the corresponding bipartite graph, $\bar{k}$: average degree of the node, $\bar{s}$: average size of the hyperedge, $\bar{c}$: average two-mode clustering coefficient of the node, $\bar{l}$: average shortest path length between nodes.

Data	N	M	$\mathcal{M}$	$\bar{k}$	$\bar{s}$	$\bar{c}$	$\bar{l}$	Refs.
email-Enron	143	1,512	4,550	31.82	3.01	0.68	2.08	[7, 3]
NDC-classes	628	816	5,688	9.06	6.97	0.31	3.53	[3]
primary-school	242	12,704	30,729	126.98	2.42	0.70	1.73	[16, 3]

the size distributions of the hyperedge in the subgraphs of size $d_e \in \{0, 1\}$ or less in the given bipartite graph [10]. We extend the model to the case of $d_v = 2.5+$ and $d_e \in \{0, 1\}$ to generate a randomized bipartite graph that preserves the joint degree distribution and the degree-dependent two-mode clustering coefficient of the node in addition to the average or distribution of the hyperedge's size.

In the model with $d_v = 2.5+$ and $d_e \in \{0, 1\}$, we first generate a randomized bipartite graph with $d_e = 2$ and given d_e using the original hyper dK-series. Then, we repeat the rewiring process for the generated bipartite graph [10]. We select a pair of edges, (v_i, e_j) and $(v_{i'}, e_{j'})$, in the bipartite graph such that $i \neq i'$, $j \neq j'$, and $k_i = k'_i$ uniformly at random. We replace (v_i, e_j) and $(v_{i'}, e_{j'})$ by $(v_i, e_{j'})$ and $(v_{i'}, e_j)$ if and only if the normalized L_1 distance defined as

$$D_{2.5+} = \frac{\sum_{k=1}^{M} |c'(k) - c(k)|}{\sum_{k=1}^{M} c(k)}$$

decreases, where $c'(k)$ represents the degree-dependent two-mode clustering coefficient of the node for the hypergraph after rewiring the edge-pair. The rewiring procedure preserves the pairwise joint degree distribution of the node and the size of each hyperedge. We repeat the rewiring attempts $R = 500\mathcal{M}$ times.

4 Experiments

We apply the extended hyper dK-series to three empirical hypergraphs. The email-Enron hypergraph is an email network [7, 3], where nodes are email addresses and hyperedges are sets of all addressees of senders and receivers of each email. The NDC-classes hypergraph is a drug network [3], where nodes are class labels and hyperedges are sets of class labels applied to each drug. The primary-school hypergraph is a contact network [16, 3], where nodes are people and hyperedges are sets of people who contact each other face-to-face. Table 1 shows the properties of the largest connected component for the data sets.

Table 2 shows the distance in five properties between the original hypergraph and the hypergraphs generated by the hyper dK-series with $d_v \in \{0, 1, 2, 2.5, 2.5+\}$ and $d_e \in \{0, 1\}$. We calculated the Kolmogorov-Smirnov distance between the cumulative distributions of the degree distribution of the node for the original hypergraph and the generated hypergraphs. For the other properties, we calculated

Table 2. Distance between the empirical hypergraphs and those generated by the reference models.

Data	(d_v, d_e)	$P(k)$	$k_{\mathrm{nn}}(k)$	$c(k)$	$P(s)$	$P(l)$
email-Enron	$(0,0)$	0.434	0.420	0.781	0.158	0.595
	$(1,0)$	0.000	0.202	0.206	0.160	0.491
	$(2,0)$	0.000	0.012	0.207	0.160	0.429
	$(2.5,0)$	0.000	0.013	0.222	0.160	0.416
	$(2.5+,0)$	0.000	0.013	0.023	0.160	0.368
	$(0,1)$	0.406	0.412	0.772	0.000	0.647
	$(1,1)$	0.000	0.200	0.197	0.000	0.491
	$(2,1)$	0.000	0.035	0.191	0.000	0.452
	$(2.5,1)$	0.000	0.027	0.198	0.000	0.427
	$(2.5+,1)$	0.000	0.032	0.026	0.000	0.414
NDC-classes	$(0,0)$	0.614	0.741	0.962	0.252	1.585
	$(1,0)$	0.000	0.388	0.368	0.248	1.322
	$(2,0)$	0.000	0.046	0.208	0.248	0.799
	$(2.5,0)$	0.000	0.045	0.201	0.248	0.712
	$(2.5+,0)$	0.000	0.043	0.035	0.248	0.467
	$(0,1)$	0.597	0.751	0.951	0.000	1.612
	$(1,1)$	0.000	0.389	0.328	0.000	1.417
	$(2,1)$	0.000	0.022	0.158	0.000	0.749
	$(2.5,1)$	0.000	0.019	0.156	0.000	0.722
	$(2.5+,1)$	0.000	0.021	0.023	0.000	0.609
primary-school	$(0,0)$	0.380	0.429	0.848	0.303	0.856
	$(1,0)$	0.000	0.089	0.121	0.307	0.707
	$(2,0)$	0.000	0.006	0.111	0.307	0.374
	$(2.5,0)$	0.000	0.005	0.109	0.307	0.371
	$(2.5+,0)$	0.000	0.007	0.008	0.307	0.313
	$(0,1)$	0.368	0.451	0.868	0.000	0.535
	$(1,1)$	0.000	0.089	0.126	0.000	0.434
	$(2,1)$	0.000	0.014	0.166	0.000	0.246
	$(2.5,1)$	0.000	0.015	0.180	0.000	0.218
	$(2.5+,1)$	0.000	0.014	0.010	0.000	0.276

the normalized L^1 distance between the properties for the original hypergraph and the generated hypergraphs.

We make the following observations for the three empirical hypergraphs. First, the distances for $P(k)$ is equal to 0 for $d_v = 2.5+$ and $d_e \in \{0, 1\}$, as expected. Second, the distances for $k_{\mathrm{nn}}(k)$ are quite small values in the models with $d_v \in \{2.5, 2.5+\}$ and $d_e \in \{0, 1\}$, which indicates that the models approximately preserve $k_{\mathrm{nn}}(k)$. Third, the distance for $c(k)$ is much smaller in the model with $d_v = 2.5+$ than that in the model with $d_v = 2.5$ for any $d_e \in \{0, 1\}$. Fourth, the model with $(d_v, d_e) = (2.5+, 0)$ and $d_e = 0$ has a somewhat distance for $P(s)$ but that with $(d_v, d_e) = (2.5+, 1)$ has no distance for $P(s)$, as expected. Finally, the model with $d_v = 2.5+$ often preserves more accurately $P(l)$ than that with $d_v = 2.5$ for any $d_e = \{0, 1\}$, whereas the distance is still not small in

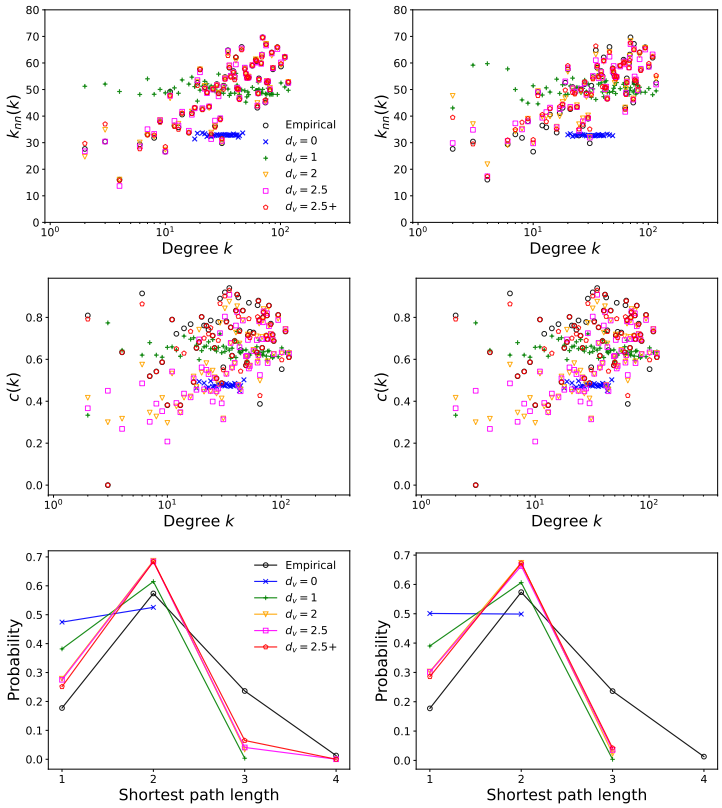

Fig. 1. Structural properties of the email-Enron hypergraph and the hypergraphs generated by the reference models. The figures in the left column show the results for $d_e = 0$. The figures in the right column show the results for $d_e = 1$.

both models. Figure 1 shows $k_{\mathrm{nn}}(k)$, $c(k)$, and $P(l)$ for the original email-Enron hypergraph and hypergraphs generated by the hyper dK-series.

5 Conclusion

We extended the hyper dK-series to generate a randomized hypergraph that preserves the pairwise joint degree distribution and the two-mode clustering coefficient. We applied the extended hyper dK-series to three empirical hypergraphs. We numerically showed that the model with $d_v = 2.5+$ preserves exactly the degree distribution of the node, approximately the pairwise joint degree distribution of the node, and approximately the degree-dependent two-mode clustering coefficient. We also found that the model with $d_v = 2.5+$ often preserves more accurately the distribution of the shortest-path length than the existing model with $d_v = 2.5$. Future work includes the application of the extended hyper dK-series to simulations of dynamical processes in hypergraphs [1].

References

1. Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., Murray, M.M., Peixoto, T.P., Vaccarino, F., Petri, G.: The physics of higher-order interactions in complex systems. Nature Physics **17**, 1093–1098 (2021)
2. Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.G., Petri, G.: Networks beyond pairwise interactions: Structure and dynamics. Physics Reports **874**, 1–92 (2020)
3. Benson, A.R., Abebe, R., Schaub, M.T., Jadbabaie, A., Kleinberg, J.: Simplicial closure and higher-order link prediction. Proceedings of the National Academy of Sciences **115**, E11221–E11230 (2018)
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports **424**, 175–308 (2006)
5. Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., Caldarelli, G.: The statistical physics of real-world networks. Nature Reviews Physics **1**, 58–71 (2019)
6. Gjoka, M., Kurant, M., Markopoulou, A.: 2.5K-graphs: From sampling to generation. In: 2013 Proceedings IEEE INFOCOM. pp. 1968–1976 (2013)
7. Klimt, B., Yang, Y.: The Enron corpus: A new dataset for email classification research. In: Proceedings of the 15th European Conference on Machine Learning. pp. 217–226 (2004)
8. Mahadevan, P., Krioukov, D., Fall, K., Vahdat, A.: Systematic topology analysis and generation using degree correlations. In: Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. pp. 135—-146 (2006)
9. Miyashita, R., Nakajima, K., Fukuda, M., Shudo, K.: Randomizing hypergraphs preserving two-mode clustering coefficient. In: 2023 IEEE International Conference on Big Data and Smart Computing (BigComp). pp. 316–317 (2023)
10. Nakajima, K., Shudo, K., Masuda, N.: Randomizing hypergraphs preserving degree correlation and local clustering. IEEE Transactions on Network Science and Engineering **9**, 1139–1153 (2022)
11. Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences **98**, 404–409 (2001)
12. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Physical Review E **64**, 026118 (2001)
13. Opsahl, T.: Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social Networks **35**, 159–167 (2013)
14. Orsini, C., Dankulov, M.M., Colomer-de Simón, P., Jamakovic, A., Mahadevan, P., Vahdat, A., Bassler, K.E., Toroczkai, Z., Boguñá, M., Caldarelli, G., Fortunato, S., Krioukov, D.: Quantifying randomness in real networks. Nature Communications **6**, 8627 (2015)
15. Patania, A., Petri, G., Vaccarino, F.: The shape of collaborations. EPJ Data Science **6**, 18 (2017)
16. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., Vanhems, P.: High-resolution measurements of face-to-face contact patterns in a primary school. PLOS ONE **6**, e23176 (2011)