クラスタ係数を保存するランダムハイパーグラフの生成手法

宮下　陸矢†　　中嶋　一貴†　　福田　萌斐†　　首藤　一幸†

† 東京工業大学 情報理工学院 数理・計算科学系

あらまし　ハイパーグラフは 2 個以上のノード間の相互作用を捉えるデータ構造であり，ハイパーグラフの実データはますます増えている．また，特定の統計量を保存したランダム・ハイパーグラフは，元のハイパーグラフと比較することにより，その統計量が与える影響を調べる上で有用である．本研究では，同時次数分布と，ノードの三項関係を定量化する指標であるクラスタ係数を，同時に保存するランダム・ハイパーグラフの生成手法を提案する．

キーワード　ハイパーグラフ，クラスタ係数，生成モデル

1 はじめに

あるグラフの指定した統計量は保存しつつもランダムに生成するランダム・グラフは，実世界のネットワークの構造やダイナミクスを解析する上で大きな役割を持つ．なぜなら，特定の統計量を保存するランダム・グラフと，元のグラフを比較することにより，ランダム・グラフにおいて保存されていない特性が，ネットワークにどのような影響を与えているのかを調査することができるからである．

ネットワークを表現する方法として一般的であるグラフは，エッジにより 2 つのノード間の相互作用を表しているが，実世界のネットワークには，ノードが集団間の相互作用を持つものも存在する．例えば，電子メールの送受信者のネットワーク [3] では，1 通のメールに対して複数の送受信者が存在する．また，論文の共著者のネットワーク [6], [7] では，1 つの論文に対する共著者は 2 人より多く存在し得る．このようなネットワークを表すものとして，グラフの拡張であるハイパーグラフが広く知られている．ハイパーグラフは，ノードの集合とハイパーエッジの集合から構成されており，ハイパーエッジは任意の数のノードを含む．ハイパーグラフにおいてもグラフと同様に，ランダム・ハイパーグラフは比較による構造の解析において有用である．そのため，ネットワークの構造に影響を与え得る特性を保存したランダム・ハイパーグラフが望まれる．

本研究では，ハイパーグラフにおける，ノードとハイパーエッジの指定した統計量を保存してランダム化する生成モデルである hyper dK-series [5] において，2.5K 統計量である冗長係数に代えて，2.5+K 統計量としてクラスタ係数を指定することで，クラスタ係数を保存するランダム・ハイパーグラフの生成を行う．

2 関連研究

グラフの生成モデルである dK-series [2], [4], [9] は，次数に基づいた統計量を保存するものであり，各ノードの次数に加え，次数相関やクラスタ係数を保存することができる．

また，ハイパーグラフの生成モデルの研究は多くなされている．ハイパーグラフと，ハイパーグラフを表現することができる構造である二部グラフにおいて，ノードの次数やハイパーエッジを保存するモデルが，いくつかの研究において提案されている [10], [11]．そして，中嶋ら [5] はノードの次数とハイパーエッジのサイズの厳密な保存に加え，同時次数分布と冗長係数を近似的に保存する，hyper dK-series を提案している．

本研究では，hyper dK-series において，冗長係数に代えてクラスタ係数を保存することで，ノードの次数とハイパーエッジのサイズの厳密な保存に加え，同時次数分布とクラスタ係数を近似的に保存する手法を提案する．

3 準　　備

3.1 表　　記

本論文では，ノード集合 $V = \{v_1, \ldots, v_N\}$ (N はノード数) と，ハイパーエッジ集合 $E = \{e_1, \ldots, e_M\}$ (M はハイパーエッジ数) からなる重みのないハイパーグラフを，二部グラフ $G = (V, E, \mathcal{E})$ として表現する．V, E は二部グラフの 2 つのノード集合であり，$\mathcal{E}$ は二部グラフのエッジの集合である．ここで，二部グラフ中の二つのノード v_i, e_j を接続するエッジ (v_i, e_j) が存在することと，ノード v_i がハイパーエッジ e_j に含まれることとは同値である．また，提案手法により生成するハイパーグラフの元となるハイパーグラフには，重複するハイパーエッジは含まれていないとする．

3.2 統　計　量

本節では，提案手法によって保存される二部グラフ G の統計量について述べる．まず，ハイパーグラフ G の接続行列を $B = (B_{ij})_{N \times M}$ として表す．接続行列 B の各成分は，

$$B_{ij} = \begin{cases} 1 & \text{if } (v_i, e_j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

として与えられる．接続行列 B を用いることにより，ノード v_i の次数 k_i は $k_i = \sum_{j=1}^{M} B_{ij}$，ハイパーエッジ e_j のサイズ s_j は $s_j = \sum_{i=1}^{N} B_{ij}$ として表すことができる．また，ハイパーエッジの平均サイズを $\bar{s}$，ノードの平均次数を $\bar{k}$ と表記する．

2 ノード間の同時次数分布は，[5] と同様に，[4], [9] のグラフ

上の定義を拡張し，少なくとも一つのハイパーエッジを共有するものとして定義する．すなわち，$m(k,k')$ を，次数 k のノードと次数 k' のノードが共有するハイパーエッジの数とすると，同時次数分布 $P(k,k')$ は，

$$P(k,k') = \frac{2m(k,k')}{\sum_{j=1}^{M} s_j(s_j-1)}$$

として表される．また，[5] と同様に，次数 k のノードの隣接ノードの平均次数を，[12], [13] のグラフ上の定義の拡張として，

$$k_{nn}(k) = \frac{\sum_{k'=1}^{M} k'P(k,k')}{\sum_{k'=1}^{M} P(k,k')}$$

と定義する．

二部グラフの各ノードのクラスタ係数 $c(v_i)$ は，[8] の定義を用いて，以下の式のように表す．

$$c(v_i) = \frac{(\text{ノード } v_i \text{ を中心とする閉じた 4path の数})}{(\text{ノード } v_i \text{ を中心とする 4path の数})}$$

すなわち，ノード v_i のクラスタ係数は，v_i を中心とする長さ 4 のパスの数のうち，パスの最初と最後のノードが共通のノードと接続しているものの割合として表される．また，ノードの平均クラスタ係数を $\bar{c}$ と表記する．

3.3 hyper dK-series

グラフの生成モデルである dK-series [2], [4], [9] をハイパーグラフに拡張した hyper dK-series [5] は，ノードの同時次数分布をパラメータ $d_v = \{0,1,2,2.5\}$，ハイパーエッジのサイズ分布をパラメータ $d_e = \{0,1\}$ の範囲内で保存する．ここで，$d_v = 0$ はノードの平均次数，$d_v = 1$ は各ノードの次数をそれぞれ保存し，$d_v = 2$ はノードの同時次数分布を近似的に保存する．また，$d_e = 0$ はハイパーエッジの平均サイズ，$d_e = 1$ は各ハイパーエッジのサイズをそれぞれ保存する．$d_v = 2.5$ は，$d_v = 2$ から $d_v = 3$ の間の同時次数分布を意味しており，冗長係数が指定されている．本研究では，$d_v = 2.5+$ として，次数依存のクラスタ係数 $\bar{c}(k) = \frac{1}{N(k)} \sum_{i=1,k_i=k}^{N} c(v_i)$ を指定する．

4 提案手法

本章では，クラスタ係数を保存するランダム・ハイパーグラフの生成手法を述べる．

提案手法は，入力としてハイパーグラフ $G = (V,E)$ と，$d_v = \{0,1,2,2.5+\}, d_e = \{0,1\}$ の範囲の値をとるパラメータ (d_v, d_e) を受け取り，生成ハイパーグラフ $\tilde{G} = (\tilde{V}, \tilde{E})$ を返す．

パラメータ (d_v, d_e) が $d_v = \{0,1,2\}, d_e = \{0,1\}$ の範囲内の値をとるとき，hyper dK-series [5] と同様にハイパーグラフの生成を行う．$(d_v, d_e) = (2.5+, 0)$ または $(d_v, d_e) = (2.5+, 1)$ のとき，それぞれ $(d_v, d_e) = (2,0)$，$(d_v, d_e) = (2,1)$ のランダム・ハイパーグラフに対して以下の再配線を行うことにより，次数依存のクラスタ係数を近似的に保存する．

まず，$i \neq i', j \neq j', k_i = k_i'$ となるエッジのペア $(v_i, e_j), (v_{i'}, e_{j'})$ を一様にランダムに選択する．そして，$(v_i, e_j), (v_{i'}, e_{j'})$ を $(v_i, e_{j'}), (v_{i'}, e_j)$ に置き換えたとき，以下のように定義される距離

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

が減少するのならば，実際にエッジの張り替えを行う．ここで，$c(k), c'(k)$ はそれぞれ，元のハイパーグラフ，ランダム・ハイパーグラフの次数依存のクラスタ係数である．このエッジの張り替え手法は，hyper dK-series [5] において 2.5K 統計量の冗長係数を合わせる手法と同様のものであり，張り替えの前後で，$d_v = 2$ 以下の統計量である，各ノードの次数，各ハイパーエッジのサイズ，同時次数分布を保存する．

表 1　データセット

データ	N	M	$\mathcal{M}$	$\bar{k}$	$\bar{s}$	$\bar{c}$	$\bar{l}$	参考文献
email-Enron	143	1512	4550	31.82	3.01	0.68	2.08	[1], [3]
NDC-classes	628	816	5688	9.06	6.97	0.31	3.53	[1]
contact-primary-school	242	12704	30729	126.98	2.42	0.70	1.73	[1], [14]

5 実　　験

5.1 データセット

実世界のハイパーグラフに対し，クラスタ係数を保存する hyper dK-series を適用する．email-Enron ハイパーグラフは電子メールのネットワーク [1], [3] であり，電子メールのアドレスをノードとし，それぞれの電子メールの送受信者全体の集合をハイパーエッジとしている．NDC-classes ハイパーグラフは医薬品のネットワーク [1] であり，クラスラベルをノード，ある医薬品に適用されているクラスラベルの集合をハイパーエッジとしている．contact-primary-school ハイパーグラフは小学校での人々の接触のネットワーク [1], [14] であり，人間をノード，接触した人々の集合をハイパーエッジとしている．用いたデータセットに対し，重複するハイパーエッジを取り除き，最大連結成分を抽出する前処理を行った．表 1 に前処理を行なったデータセットの統計量を示す．表中の $\bar{l}$ は，平均最短経路長を表している．

5.2 生成結果

それぞれのデータセットに対して，提案手法によりランダム・ハイパーグラフの生成を行ったところ，表 2 に示す結果が得られた．なお，$d_v = 2.5+$ では，[5] の $d_v = 2, 2.5$ と同様に，エッジの張り替えの試行を $R = 500\mathcal{M}$ 回行った．

表 2 は，生成ハイパーグラフと元のハイパーグラフの統計量の誤差を表している．$\Delta P(k)$ は次数分布の累積分布のコルモゴロフ-スミルノフ距離である．$\Delta k_{nn}(k)$ は次数 k を持つノードの隣接ノードの平均次数，$\Delta c(k)$ は次数依存のクラスタ係数，$\Delta P(l)$ は最短経路長の分布の正規化された距離を表している．

また，元のハイパーグラフと次数分布が一致する $d_v = 1, 2, 2.5+$ の生成ハイパーグラフにおける，元のハイパーグラフとの各次数の隣接ノードの平均次数 $k_{nn}(k)$，クラスタ係数 $c(k)$ の誤差と，元のハイパーグラフと生成ハイパーグラフの最短経路長分布を，次の図 1 に示す．

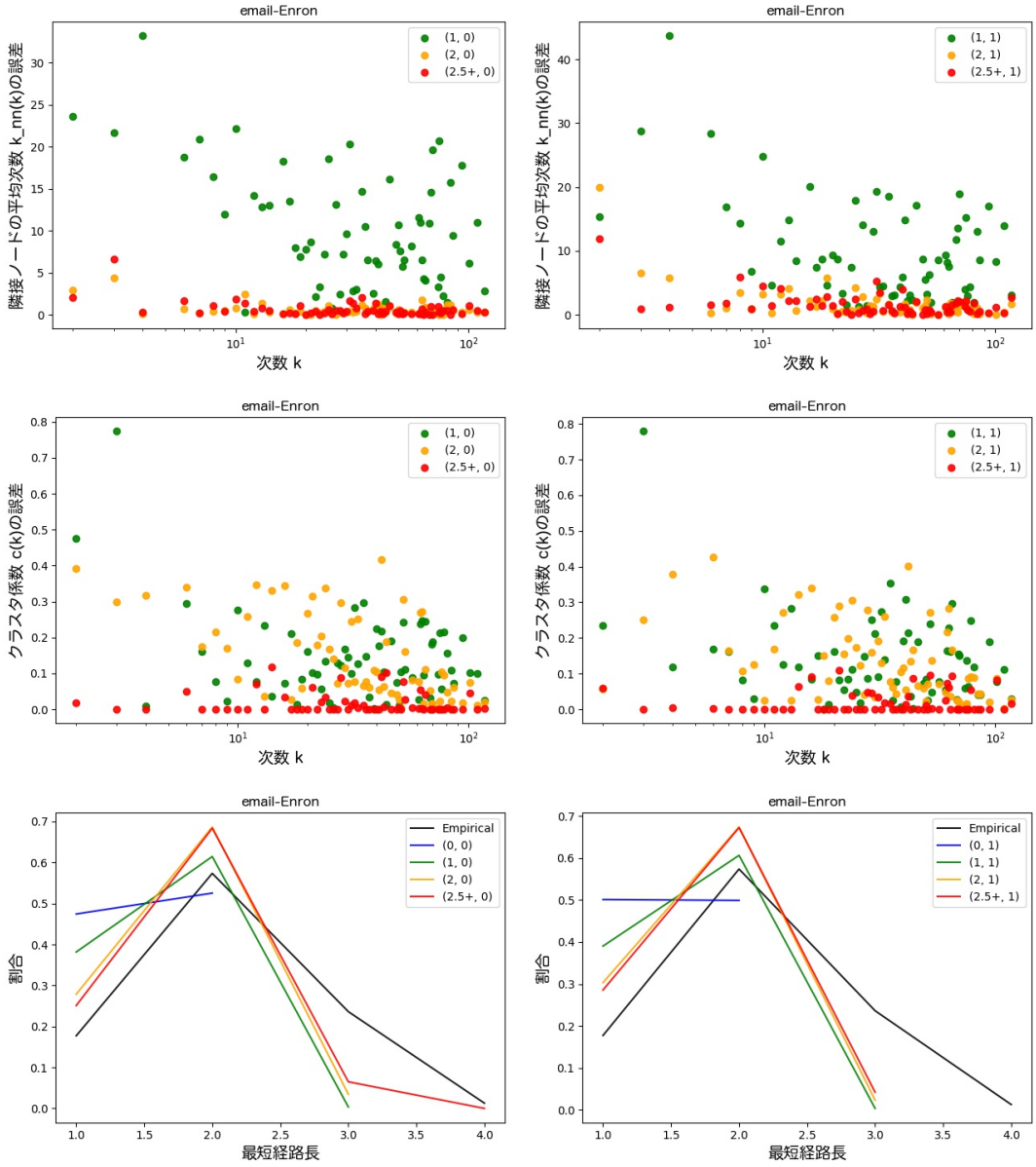

図 1　生成ハイパーグラフの統計量の誤差

5.3　考　　察

いずれのデータセットから生成したランダム・ハイパーグラフも，$(d_v, d_e) = (2.5+, 0), (2.5+, 1)$ では共に，表 2 中の次数分布の誤差 $\Delta P(k)$ は 0 であり，$d_v = 0, 1$ の指定する統計量である平均次数，各ノードの次数を厳密に保存している．また，次数 k を持つノードの隣接ノードの平均次数 $k_{nn}(k)$ の誤差 $\Delta k_{nn}(k)$ は，表 2 において $(d_v, d_e) = (2, 0), (2, 1)$ の生成ハイパーグラフと同様に小さい値をとっていることから，$d_v = 2$ の指定する統計量である同時次数分布も近似的に保存されている．そして，提案手法において保存することを目的としている，$d_v = 2.5+$ の指定するクラスタ係数をみると，いずれのデータセットにおいても，表 2 中の誤差 $\Delta c(k)$ は，$d_v = 2$ 以下の生成ハイパーグラフと比較して小さい値となっており，図 1 に示した email-Enron ハイパーグラフの各次数での誤差は，$d_e = 0, 1$ のどちらにおいても多くの次数で 0 に近い値をとっている．このことから，実際にクラスタ係数の近似的な保存がされていると考えられる．これらの結果から，提案手法は，同時次数分布とクラスタ係数を保存するハイパーグラフのランダム化を実現した．

さらに，保存を意図していない最短経路長分布の誤差 $\Delta P(l)$ は，いずれの生成ハイパーグラフにおいても大きく，正確な近似はされていないが，$d_v = 2.5+$ の生成ハイパーグラフが他の生成ハイパーグラフに対して比較的小さいという結果になった．また，いずれのデータセットにおいても $d_v = 2.5+$ の生成ハイパーグラフの平均最短経路長は元のハイパーグラフのものより小さい．この結果は，[5] の $d_v = 2.5$ の結果と同様である．このことから，最短経路長を増加させる要因とされるコミュニティ構造は，$d_v = 2.5+$ においても保存されていないと考えられる．

表 2　生成ハイパーグラフの誤差

データ	(d_v, d_e)	$\Delta P(k)$	$\Delta k_{nn}(k)$	$\Delta c(k)$	$\Delta P(l)$
email-Enron	$(0,0)$	0.434	0.420	0.781	0.595
	$(1,0)$	0.000	0.202	0.206	0.491
	$(2,0)$	0.000	0.012	0.207	0.429
	$(2.5+,0)$	0.000	0.013	0.023	0.368
	$(0,1)$	0.406	0.412	0.772	0.647
	$(1,1)$	0.000	0.200	0.197	0.491
	$(2,1)$	0.000	0.035	0.191	0.452
	$(2.5+,1)$	0.000	0.032	0.026	0.414
NDC-classes	$(0,0)$	0.614	0.741	0.962	1.585
	$(1,0)$	0.000	0.388	0.368	1.322
	$(2,0)$	0.000	0.046	0.208	0.799
	$(2.5+,0)$	0.000	0.043	0.035	0.467
	$(0,1)$	0.597	0.751	0.951	1.612
	$(1,1)$	0.000	0.389	0.328	1.417
	$(2,1)$	0.000	0.022	0.158	0.749
	$(2.5+,1)$	0.000	0.021	0.023	0.609
contact-primary-school	$(0,0)$	0.380	0.429	0.848	0.856
	$(1,0)$	0.000	0.089	0.121	0.707
	$(2,0)$	0.000	0.006	0.111	0.374
	$(2.5+,0)$	0.000	0.007	0.008	0.313
	$(0,1)$	0.368	0.451	0.868	0.535
	$(1,1)$	0.000	0.089	0.126	0.434
	$(2,1)$	0.000	0.014	0.166	0.246
	$(2.5+,1)$	0.000	0.014	0.010	0.276

6 結　　論

本研究では，ハイパーグラフの生成モデルである hyper dK-series [5] において，ノードの 2.5+K 統計量としてクラスタ係数を指定することで，hyper dK-series に基づく，同時次数分布とクラスタ係数を保存するハイパーグラフのランダム化手法を提案した．本研究では，実世界のハイパーグラフに対して，提案手法によるランダム化を行なった．3 つの実世界のデータセットに対して $(d_v, d_e) = (2.5+,0), (2.5+,1)$ としてランダム・ハイパーグラフの生成を行なったところ，いずれのデータセットにおいても，実際に $d_v = 0, 1$ の指定するノードの平均次数，各ノードの次数を厳密に保存し，$d_v = 2, 2.5+$ の指定する同時次数分布，クラスタ係数を近似的に保存するランダム・ハイパーグラフが得られた．また，最短経路長分布は，$d_v = 2.5+$ においても正確に近似されてはいないが，いずれのデータセットにおいても，$d_v = 2.5+$ の生成ハイパーグラフは，他の生成ハイパーグラフに対して，比較的小さい誤差となった．

文　　献

[1] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, “Simplicial closure and higher-order link prediction,” *Proc. Natl. Acad. Sci. USA*, vol. 115, pp. E11 221-E11 230, 2018.

[2] M. Gjoka, M. Kurant, and A. Markopoulou, “2.5K-graphs: From sampling to generation,” in *2013 Proceedings of IEEE INFOCOM*, 2013, pp. 1968-1976.

[3] B. Klimt and Y. Yang, “The Enron corpus: A new dataset for email classification research,” in *European Conference on Machine Learning*, 2004, pp. 217-226.

[4] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, “Systematic topology analysis and generation using degree correlations,” *SIGCOMM Comput. Commun. Rev.*, vol. 36, pp. 135-146, 2006.

[5] K. Nakajima, K. Shudo, and N. Masuda, ”Randomizing Hypergraphs Preserving Degree Correlation and Local Clustering,” *IEEE Transactions on Network Science and Engineering*, 2021.

[6] M. E. J. Newman, “The structure of scientific collaboration networks,” *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 404-409, 2001.

[7] A. Patania, G. Petri, and F. Vaccarino, “The shape of collaborations,” *EPJ Data Sci.*, vol. 6, 2017, Art. no. 18.

[8] T. Opsahl, “Triadic closure in two-mode networks: Redefining the global and local clustering coefficients,” *Social networks*, 2013, 35.2, 159-167.

[9] C. Orsini, M. M. Dankulov, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov, “Quantifying randomness in real networks,” *Nat. Commun.*, vol. 6, 2015, Art. no. 8627.

[10] P. S. Chodrow, “Configuration models of random hypergraphs,” *J. Complex Netw.*, vol. 8, 2020, Art. no. cnaa018.

[11] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Phys. Rev. E*, vol. 64, 2001, Art. no. 026118.

[12] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, “Complex networks: Structure and dynamics,” *Phys. Rep.*, vol. 424, pp. 175-308, 2006.

[13] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, “Dynamical and correlation properties of the Internet,” *Phys. Rev. Lett.*, vol. 87, 2001, Art. no. 258701.

[14] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, “High-resolution measurements of face-to-face contact patterns in a primary school,” *PLoS ONE*, vol. 6, 2011, Art. no. e23176.