グラフ生成モデルの発展と今後の展望 ～統計的生成モデルからDeep Graph Generatorsまで～

Recent Advances and Present Challengings in Graph Generative Models:
Stochastic Models and Deep Graph Generators

渡部康平 [1]
Kohei Watabe

長岡技術科学大学 大学院工学研究科 [1]
Graduate School of Engineering, Nagaoka University of Technology

1 はじめに

グラフ構造は極めて汎用に利用可能な基本的なデータ構造であり，現実世界における様々なものがグラフにより抽象表現可能である．グラフはノード (節点，頂点) の集合とエッジ (リンク，辺) の集合で表される数学的構造であり，対象の相互関係をシンプルに表現可能である．鉄道網，バス路線図，道路網などの流通網や，インターネットや電話網などの通信ネットワークの分野では，古くから対象を数学的に取り扱うための手段としてグラフが活用されてきた．他にもソーシャルネットワークにおける人間関係や，電気工学における電気回路，分子工学における分子構造もグラフにより表現される．ツリー構造もグラフ構造の特殊ケースであり，データベースやプログラムコードの構造を表すデータ構造としても頻繁に登場する．近年目覚ましい発展を遂げている機械学習分野でも，決定木のツリー構造やニューラルネットワークのネットワーク構造もグラフによる表現である．

グラフを人工的に生成する技術は，通信ネットワーク，ソーシャルネットワーク，交通網，データベース，分子工学，疫学など様々な分野においてアプリケーションがあり重要である．各種分野におけるグラフ上のシミュレーションでは，特定の特徴を有した多数のグラフでシミュレーションを繰り返すことで初めて手法の有効性を立証できるが，実グラフのデータセットを常に十分量入手可能であるとは限らない．加えて，未来のグラフや存在しないグラフ構造を生成することで予測タスクにも適用可能であり，未来のソーシャルネットワークの構造予測や，未知の分子構造の生成による新薬開発，書きかけのプログラムコードからのサジェスト機能などにも活用ができる．

グラフの生成問題は歴史的には古くから研究されており，様々な生成モデルが提案されているが，共通するのは広大なグラフの空間から特定の特徴を備えたグラフをサンプリングする試みであるという点である．1959年に提案された Erdős-Rényi (ER) モデル [1] に始まり，スモールワールド性やスケールフリー性などの特徴の再現を目指した Watts–Strogatz (WS) モデル [2] や Barabási–Albert (BA) モデル [3] など，数多くのモデルが提案されてきた．近年では，深層学習を活用した生成技術の発展に伴い，実世界に存在するグラフのデータの特徴を学習し，再現する Deep graph generators と呼ばれるグラフ生成モデルが登場するようになってきた [4, 5]．多くの Deep graph generators が与えられたデータセットと同じ特徴のグラフを生成する中で，我々は与えられたグラフデータの特徴を多面的に再現しつつ，特定の特徴量を連続的に指定可能な GraphTune を提案してきた [6, 7]．

本稿では，統計的生成モデルから Deep Graph Generators まで代表的なグラフ生成モデルと我々が提案する GraphTune を紹介しつつ，グラフ生成モデルの今後の展開について言及する．本稿の構成は以下の通りである．まず，2 章と 3 章で代表的な統計的グラフ生成モデルと Deep graph generators をそれぞれ紹介し，4 章で我々の提案するグラフ生成モデルについて紹介する．5 章では，これまで紹介してきたグラフ生成モデルの課題と今後の発展の方向性について記述する．最後に，6 章で本稿のまとめを述べる．

2 統計的生成グラフモデル

1959 年に提案された Erdős-Rényi (ER) モデル [1] は最も原始的なグラフ生成モデルであり，与えられたノード間に一定確率でエッジを張る極めてシンプルな生成アルゴリズムを持つ．ER モデルはノード数 n とノード間にエッジを張る確率 p をパラメータに持ち，任意の 2 ノード間に確率 p でエッジを張る．ランダムにエッジを張るだけのため，ノードの次数 (ノードにつながるエッジの数) は均質になり，特別なノードが存在しないグラフが生成される．

2000 年頃には実世界のグラフが持つスモールワールド性やスケールフリー性といった特徴に着目し，それらを再現する統計的生成モデルが提案されている [2, 3]．スモールワールド性は，グラフのノード数 n に比して任意の二点間の平均最短経路長 L が小さい特徴を指しており，論文の共著関係などの人間関係のネットワークや送電網などの実グラフで共通に現れる特徴であることが報告されている．n に対して平均最短経路長 L が $\log n$ でしか増加しないとき，スモールワールド性を満たすと定義される．Watts らはスモールワールド性を，平均次数が $2K$ であるリング状格子から各エッジを確率 p で別のノードに張り替えるという非常にシンプルなアルゴリズムで再現してみせる Watts-Strogatz (WS) モデルを提案した [2]．一方，スケールフリー性は次数の偏りに関

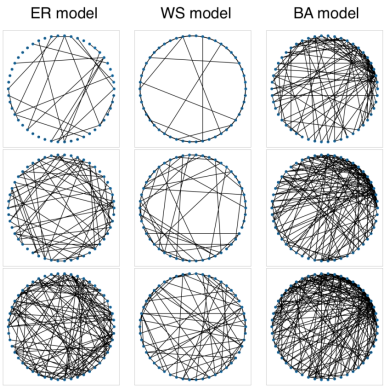

図 1　各種統計的グラフ生成モデルによる生成グラフ

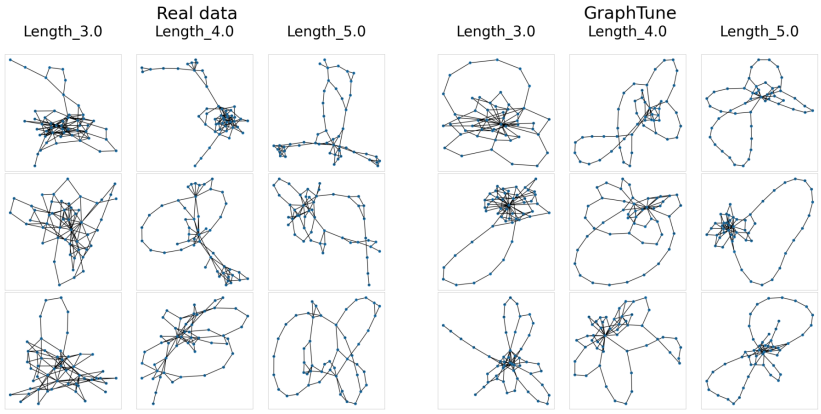

図 2　GraphTune 及び Deep Graph Generators による生成グラフ

する特徴であり，ソーシャルネットワークや食物連鎖の関係などスケールフリー性を有するグラフでは，ごく一部のノードが多数のノードとエッジでつながる一方で，多数のノードは少数のノードとしかつながっていない．より厳密には，k を次数としてノードの次数の分布 $f(k)$ が $f(k) \propto k^{-\gamma}$ $(2 \leq \gamma \leq 3)$ となっているネットワークを指す．Albert らは，m 個の完全グラフからスタートし，m 本のエッジを持つ新たなノードを追加していく Barabási-Albert (BA) モデルを提案し，$\gamma = 3$ のスケールフリーグラフの生成を可能にした．ER モデル，WS モデル，BA モデルで生成したグラフの例を図 1 に示す．

統計的生成モデルの発達とともに，前述のスモールワールド性やスケールフリー性に代表される様々な観点のグラフの特徴が注目され，それらを定量化するための特徴量が検討されてきた．上述のスモールワールド性の指標として使われる平均最短経路長やスケールフリー性を定量化した次数分布の冪指数 γ の他には，平均次数，クラスタリング係数，モジュラリティなどがよく使われる．クラスタリング係数は，ローカルなつながりの強さを表す指標であり，各ノードから伸びるエッジが三角形を構成する割合の平均により定義される．モジュラリティも同様にローカルなつながりの強さを表す指標で，Louvain アルゴリズムなどを用いてグラフをクラスタに分割した上で，クラスタ間のエッジの密度で定義される．これら以外にも直径，エッジ密度，クリーク，クラスタ数，最大コンポーネントサイズ，Assortativity，Reciprocity，隣接行列の固有値，中心性など，多数の指標が存在し，それらのすべてを忠実に再現可能な統計的生成モデルは存在しない．図 3 には，$K = 2$ とした WS モデルで生成した 300 個のグラフの特徴量を計算し，横軸及び縦軸のラベルに示す特徴量の散布図を表示しており，同時に対角線上のプロットには各特徴量の分布を表示している．図 3 のように WS モデルが生成するグラフの特徴量を多面的に描画してみると，各特徴量がモデル固有の高い依存関係にあり独立した調整はできていないことがわかる．特徴量の性質を鑑みれば，平均最短経路長を固定してもクラスタリング係数は高い自由度で決めることができるはずだが，特定の値のグラフしか生成されておらず，グラフが存在する空間からの均質なサンプリングが実現できていないことがわかる．

3　Deep Graph Generators

統計的生成モデルは実世界にあるグラフの特定の特徴を切り出し，それらをシンプルなモデルで表現することで膨大な数の研究に影響を与えて貢献したが，一方で，実世界にあるグラフが持つ多数の特徴を多面的に再現することはできない．前述のとおり実世界に存在するグラフには多数の特徴を複合的に含有し，グラフを表す特徴量はドメイン毎に異なる．WS モデルや BA モデルのように特定の特徴をパラメータで調整可能なモデルも，すべての特徴を同時に調整可能なわけではない上に，モデルにより各特徴量は依存関係にあり独立に指定可能なわけではない．

これらの問題に対して，2010 年代後半頃から GraphGen [5] に代表される Deep Graph Generators と呼ばれる深層学習による生成技術が研究されるようになってきており，グラフのデータセットの特徴を多面的に模倣したグラフの生成を目指した研究が報告されている．この背景には，2014 年に Generative Adversarial Network (GAN) が提案されるなど，深層学習による生成技術の急速な発展があり，2010 年代後半から画像生成タスクなどで先行する生成技術をより複雑なデータ構造であるグラフに展開されるようになってきている．技術的な難しさとしては，グラフという複雑なデータ構造を，ベクトルやテンソルの入力を前提として設計されている深層学習の枠組みにどのように入力するかという点であり，隣接行列を画像化する，グラフ上のウォークをシーケンスデータとして取り扱うなどのアプローチが試みられてきた．中でも，NetGAN や GraphGen [5] などのグラフ上のウォークを使ったアプローチが成功を修めている．最も成功したグラフ生成モデルの一つである GraphGen では，Depth-First Search (DFS) code と呼ばれる深さ優先探索順で並べたエッジのシーケンスに変換し，DFS code を時系列処理のための深層学習の Long Short Term Memory (LSTM) に入力する．LSTM によりシーケンスの予測タスクを解くことにより，予測されるシーケン

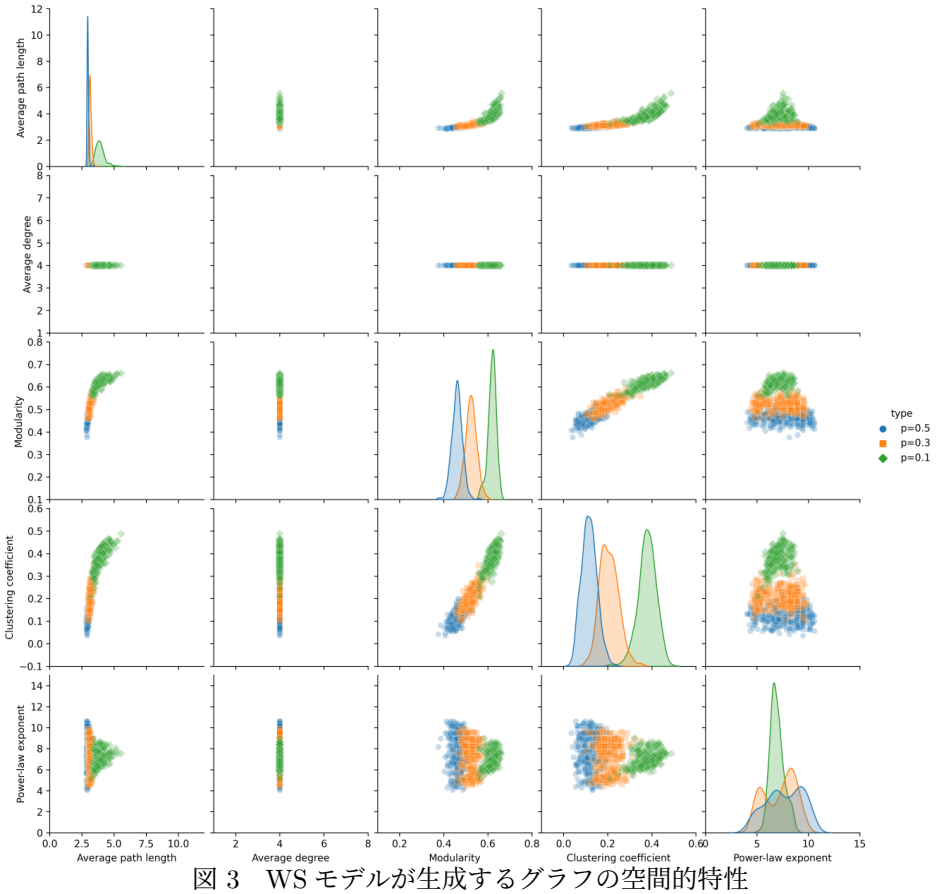

図 3　WS モデルが生成するグラフの空間的特性

スを出力してグラフ生成を実現する．

4　特徴量を調整可能なモデル

深層学習を活用した生成技術の多くは，データセットとして与えられたグラフの特徴をそのまま再現することを目標としているが，我々はグラフの特徴を多面的に保存しつつ，特定の特徴量を任意に調整可能な GraphTune [6, 7] を提案した．GraphTune では，GraphGen と同様にエッジのシーケンスによりグラフデータをニューラルネットワークに入出力するが，Variational AutoEncoder (VAE) によりグラフの特徴を潜在空間 z にエンコードしてからデコードをする構造を取ることで，乱数をもとにした多様なグラフ生成を可能にしている．GraphTune の最大の特徴は，潜在空間 z に圧縮されたグラフに関する情報のうち，特定の特徴量に関する情報を上書きすることで，データセットに含まれるグラフの多くの特徴量を保持したまま，特定の特徴量を調整したグラフを生成可能にしている点である．

GraphTune による生成グラフの検証する実験結果は，GraphTune が Deep Graph Generators のように与えられたデータセットの特徴を多面的に再現しつつ，統計的生成モデルのように注目する特徴量を明示的に指定可能であることを示している．Twitter のフォロー関係のグラフから 50 ノード分サンプリングして形成される誘導部分グラフを学習し，平均最短経路長を調整して生成したグラフの例を学習に使用したグラフとともに図 2 に示す．また，生成したグラフの各特徴量の分布を図 4 に示す．図 4 において，赤色で示されているプロットが学習に使用したデータセットの分布を示しているが，ほとんどの生成されたグラフがデータセットのグラフの分布する範囲内に収まっていることがわかる．同時に生成したグラフの平均最短経路長が指定した 3，4，5 の付近に集中して分布していることが確認でき，GraphTune が指定通りのグラフを生成していることがわかる．

5　グラフ生成モデルの今後の方向性

グラフ生成に関する研究は，統計的生成モデルから Deep Graph Generators へと発展してきたが，今後は条件付き生成に関する研究が大きく発展することが期待される．データセットとして与えられたグラフに似たグラフを生成する技術は，研究的な蓄積も増え，NetGan や GraphGen など精度の良いモデルが登場してきているが，条件付きのグラフを生成する技術については未だ十分な検討がなされていない．条件付き生成に関する研究は極めて限定的で，データセットとして与えられたグラフをグループに分けてグループごとの生成を目指した CondGen [4] がある．我々が提案する GraphTune は，条件付き生成技術の延長として特徴量に連続値として条件を加える技術を確立した．GraphTune により任意の特徴量の調整が可能になったものの，未だ条件指定の精度は高いとは言えず，複数の特徴量を任意に指定することも今の所難しい．今後は，条件指定の精度向上と複数同時指定する手法の検討が重要となると見込まれる．

条件付き生成，特に連続的に複数の統計量を指定する

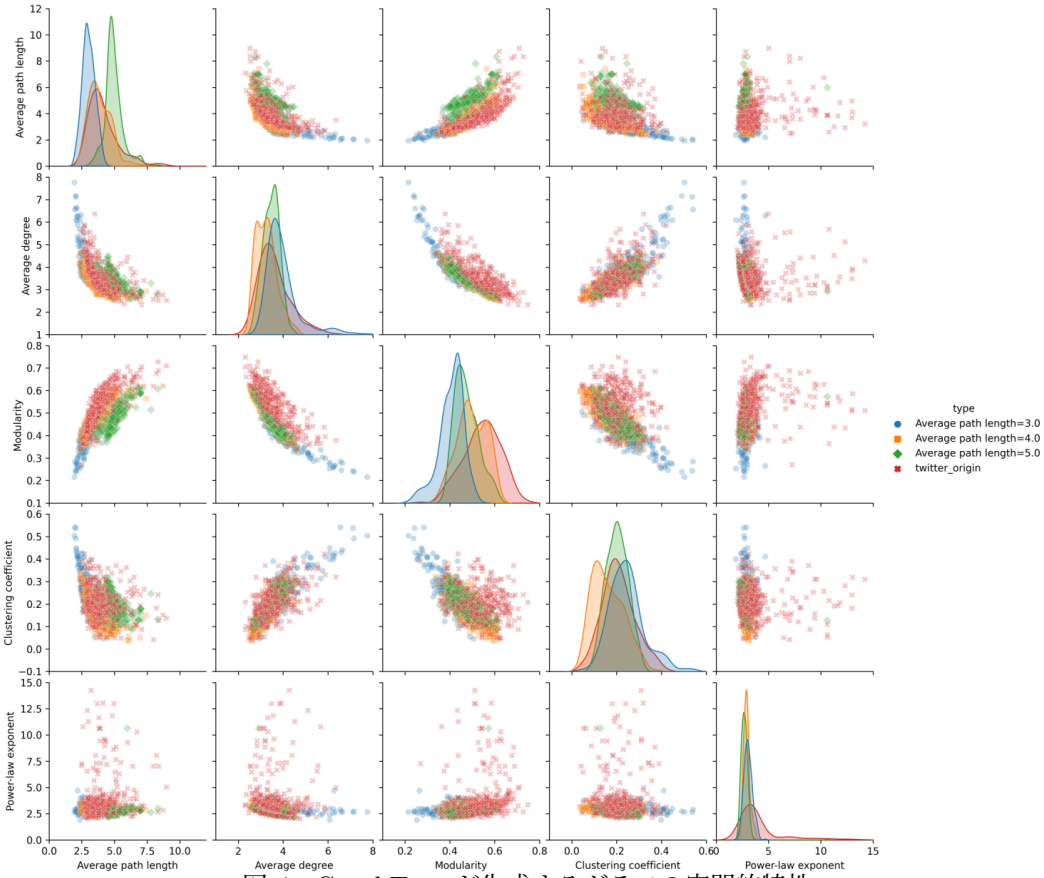

図 4　GraphTune が生成するグラフの空間的特性

技術が成熟すると，任意の特徴を持つグラフを生成可能になり，伝統的なグラフ理論によるグラフ解析に生成技術からアプローチする新たな分野が生まれる可能性がある．GraphTune の延長である複数の特徴量を任意の値に指定する技術は，広大なグラフの空間の境界に位置するグラフを生成可能である．この技術を活用すると，例えば，次数分布の平均が 4 であるグラフで最もモジュラリティが高いグラフを生成することができるため，伝統的なグラフ理論では難しかった平均次数とモジュラリティの関係を与える近似式の導出・検証などの問題に理論と生成の両面からアプローチすることが可能となる．

6　おわりに

本稿では，様々な問題に対して応用可能なグラフ生成に関する技術を古典的な統計的生成モデルと近年注目される Deep Graph Generators に分けて解説し，グラフ生成問題の今後の展望について述べた．統計的生成として ER モデルに始まり，WS モデル，BA モデルなどを紹介し，Deep Graph Generators の代表としては GraphGen と著者らが提案する特徴量の調整を可能にした GraphTune を紹介した．今後の展望としては，条件付き生成技術のより一層の発展と古典的なグラフ理論へのフィードバックについて言及した．

謝辞

本研究の一部は科研費 20H04172 の助成を受けたものである．

参考文献

[1] P. Erdös and A. Rényi, “On Random Graphs I,” *Publicationes Mathematicae*, vol. 6, no. 26, 1959.

[2] D. J. Watts and S. H. Strogatz, “Collective Dynamics of ‘Small-World’ Networks,” *Nature*, vol. 393, no. 6684, 1998.

[3] R. Albert and A.-L. Barabási, “Statistical Mechanics of Complex Networks,” *Reviews of Modern Physics*, vol. 74, no. 1, 2002.

[4] C. Yang, P. Zhuang, W. Shi, A. Luu, and P. Li, “Conditional Structure Generation through Graph Variational Generative Adversarial Nets,” in *Proc. of NeurIPS 2019*, 2019.

[5] N. Goyal, H. V. Jain, and S. Ranu, “GraphGen: A Scalable Approach to Domain-agnostic Labeled Graph Generation,” in *Proc. of WWW 2020*.

[6] S. Nakazawa, Y. Sato, K. Nakagawa, S. Tsugawa, and K. Watabe, “A Tunable Model for Graph Generation Using LSTM and Conditional VAE,” in *Proc. of ICDCS 2021 Poster Track*.

[7] K. Watabe, S. Nakazawa, Y. Sato, S. Tsugawa, and K. Nakagawa, “GraphTune: A Learning-based Graph Generative Model with Tunable Structural Features,” *CoRR*, vol. abs/2201.11494, 2022.