

INVITED TALK

生成AIと情報ネットワーク の融合:

トポロジおよびトラフィック生成の基礎理論と研究動向

埼玉大学 理工学研究科

渡部 康平

OVERVIEW

今回の発表の概要

1. 通信におけるデータ生成

3. グラフ生成モデルの発展

5. モデルの複雑性と非学習的生成

2. トラフィック生成とNIDS

4. 生成データの品質評価

PART I

通信におけるデータ生成

GENERATIVE AI TODAY

生成AIは画像と自然言語を中心に発展してきた

画像生成

拡散モデルなどにより、高精細で多様な画像を生成できるようになった

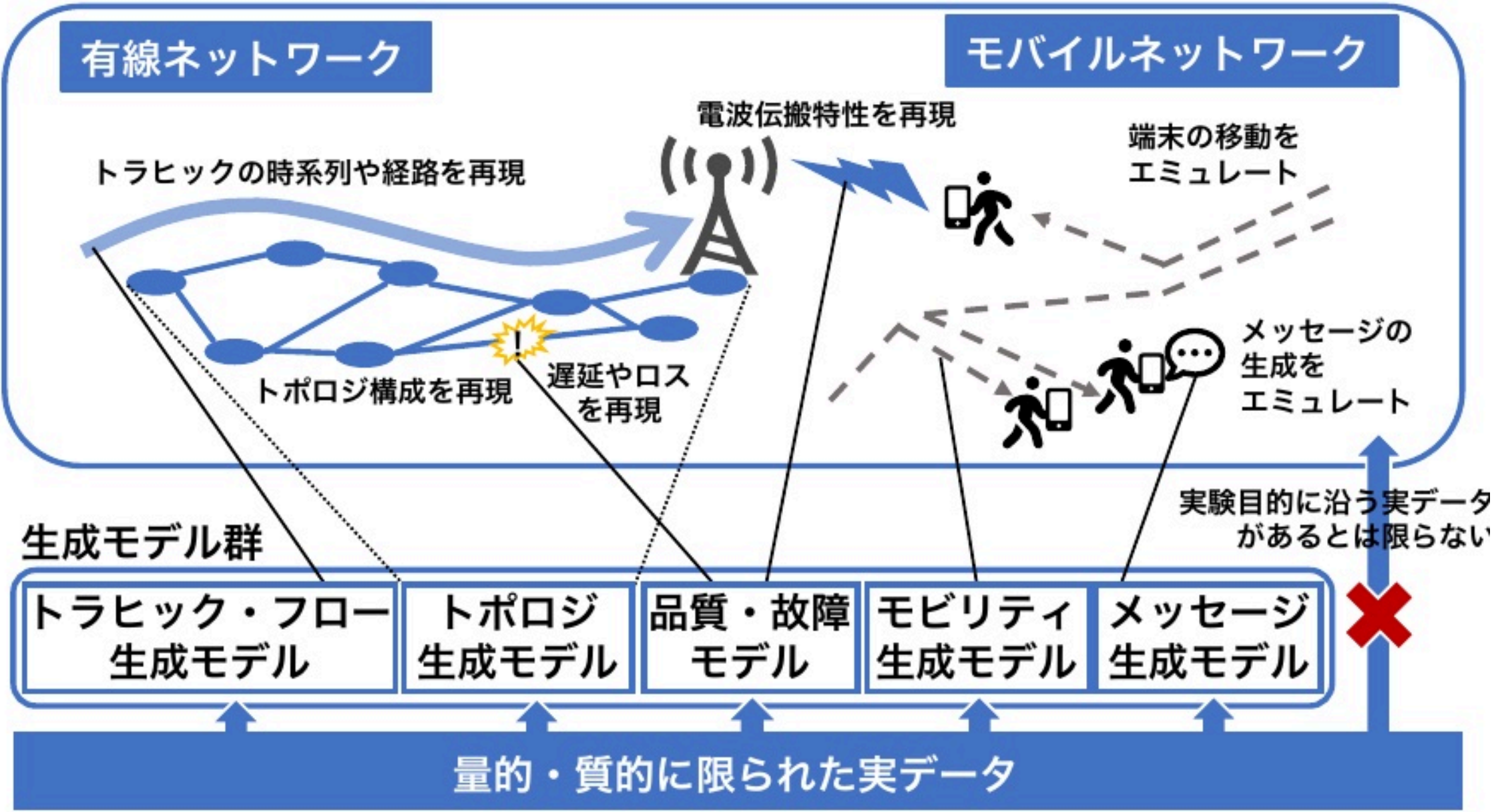
自然言語生成

大規模言語モデルにより、文脈に沿った自然な文章を生成できるようになった

では、どんなデータも同じように生成可能になったのか？ **否**
である。

COMMUNICATION DATA

通信ネットワークにおけるデータ



DATA SCARCITY

通信データを入手することの難しさ

01

ネットワーク・プロトコルの
テストやシミュレーション

説得力のある検証には
実データが不可欠

02

ところが、実データは手に入らない

非公開 通信データはプライバシーの塊

独占 一部の企業・組織が収集し、独占的に
利用

制約 公開データは限定的で、多くの研究者
が制約を受ける

GAFAによるデータ独占問題のネットワーク版

必要な特性を備え、丁度いい規模で、必要な組み合わせが揃った
実データが入手できることはほとんどない。

TWO APPROACHES

人工的なデータ生成の二つの枠組み

01

統計的生成モデル

分布や生成規則を人が設計する

- 少数のパラメータで特性を制御
- 再現対象ごとに個別設計

Poisson、MMPP、WA、BAなど

02

機械学習的生成モデル

実データから生成規則を学習する

- 複数の特性を同時に表現
- 学習データとモデル設計に依存

VAE、GAN、拡散モデルなど

本講演では、**生成モデルの発展と現状の限界・今後の展望**を捉える。

LIMITS OF STATISTICAL MODELS

統計的モデルで生成することの問題点

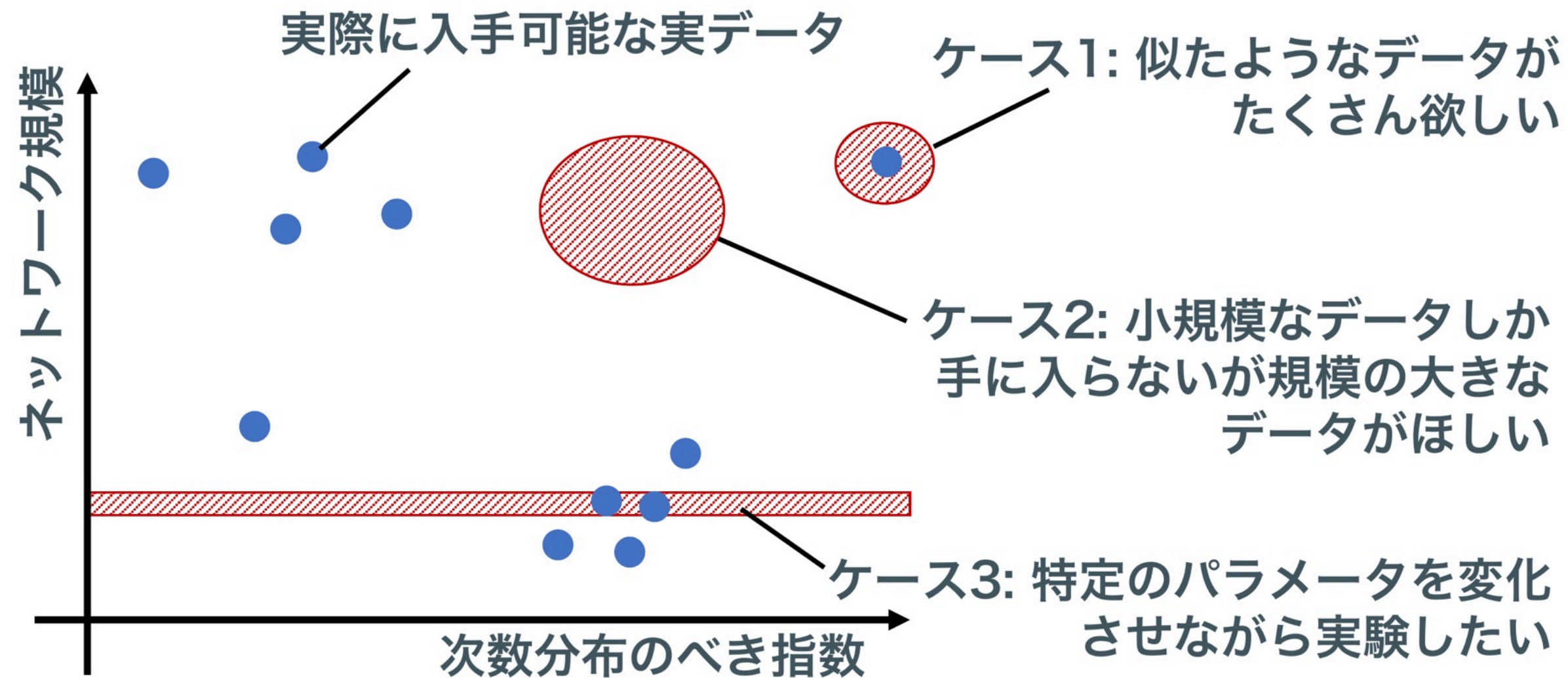


モデルが再現するのは**実データが持つ特徴の一面**。多面的な再現には専門的なチューニングが必要で、たとえチューニングしても実データによる実験結果と乖離しうる。

VALUE OF SYNTHETIC DATA

実データを模擬したデータの有用性

実データの多くの特性を維持しながら、**量的制約をなくし、必要なパラメータを調整できる。**



FEATURE-CONDITIONED GENERATION

特徴量を指定して生成することの意味

特徴量を学習データ分布の端へ向けて連続的に指定し、生成を繰り返す。

生成領域は学習範囲外の「想定外」の領域まで段階的に拡張できる。

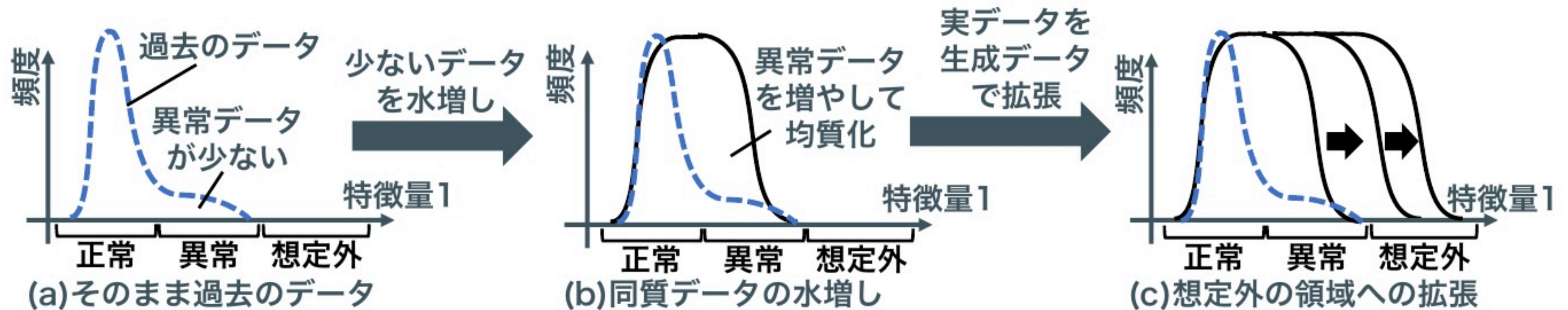


図 2: 学習データの分布とデータセット拡張技術を異常検出に適用した場合のイメージ

PART II

トラフィック生成とNIDS

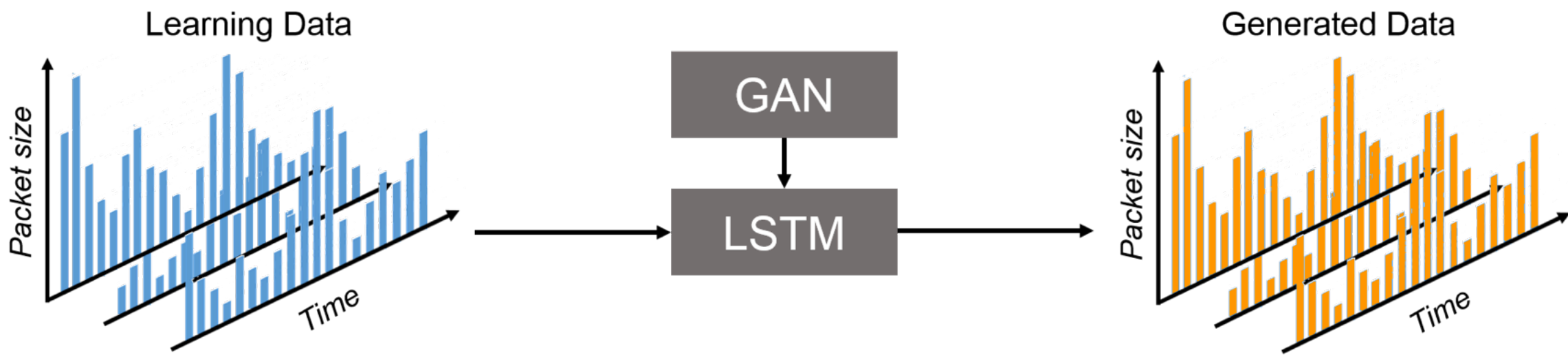
GRANULARITY

トラフィック生成の三つの粒度

バイトレベル
通信量の時系列
周期性と自己相関、多点
間相関

フローレベル
通信のまとまり
アドレス、ポート、継続
時間

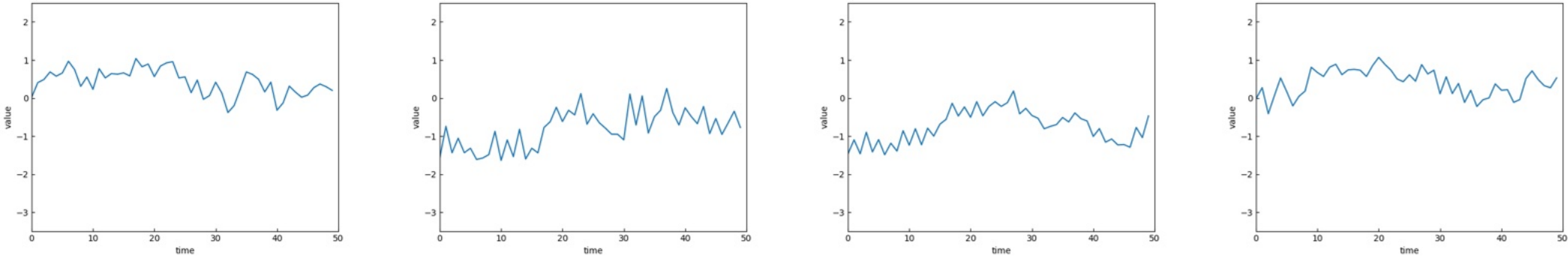
パケットレベル
通信の最小単位
順序、ヘッダ、攻撃機能



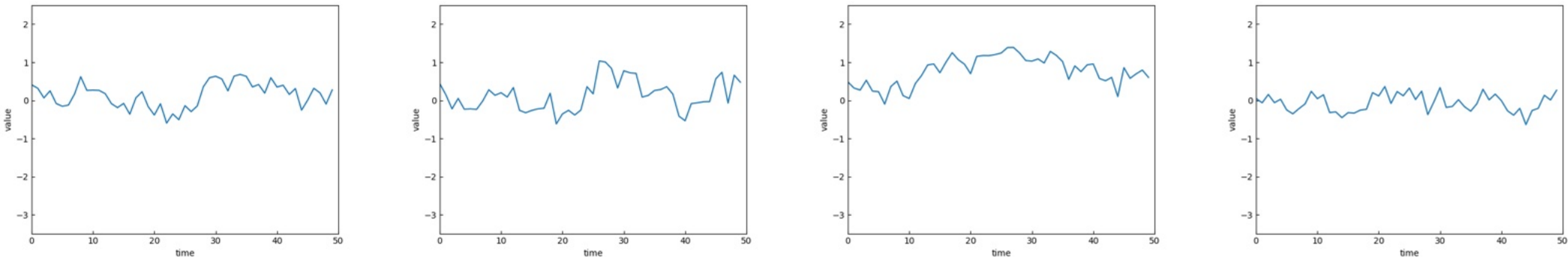
WIDE TRAFFIC DATA

学習系列と生成系列の形状

実データ



生成



RESEARCH FOCUS

攻撃トラフィック生成に関する研究

攻撃機能を保ったまま検知しにくい攻撃トラフィックを人工的に生成し、NIDSを突破させる研究を実施.



DIFFUPAC OVERVIEW

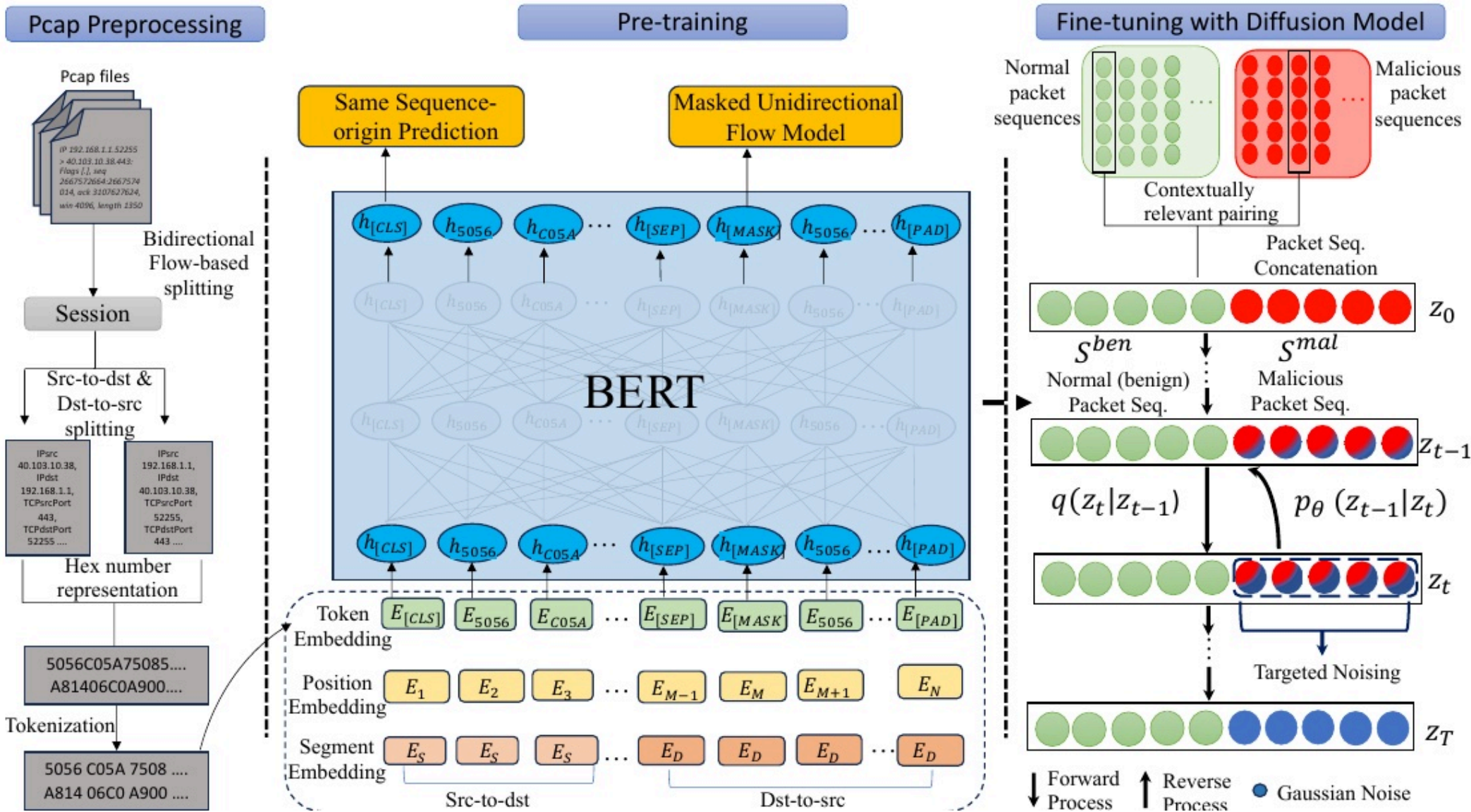
DiffuPac: 通常トラフィックを模倣した攻撃トラフィックを生成する

- | | |
|----------|--|
| 01 目的 | 攻撃機能を保ちつつ、検知に使われる特徴量を変化させ、NIDSが見逃す攻撃パターンを探索する。 |
| 02 脅威モデル | 従来法が仮定する特徴抽出器や代理分類器へのアクセスをなくし、通常通信と悪性通信の packets 列だけを利用する。 |
| 03 生成方法 | 近い通信文脈で両系列を対応付け、通常 packets 列をガイドンスとして用いる Diffusion モデルによって、悪性 packets 側だけを変化させる。 |

A. B. Jasni, A. Manada, and K. Watabe, “DiffuPac: Contextual Mimicry in Adversarial Packets Generation via Diffusion Model,” NeurIPS 2024.

DIFFUPAC

DiffuPacのアーキテクチャ



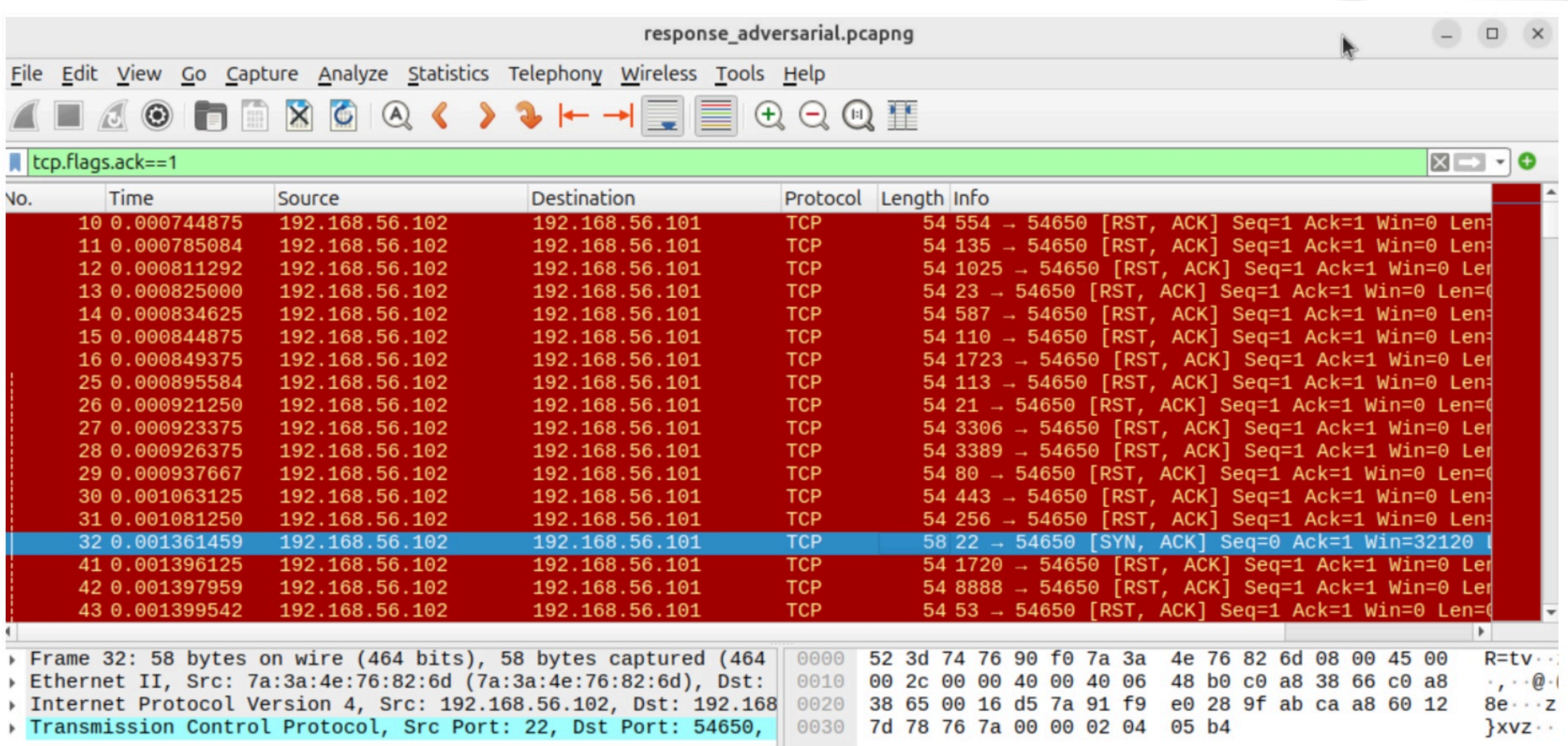
DIFFUPAC RESULTS

未知のNIDSに対して回避性能を評価

(a) Botnet								(b) MITM							
Feature Extractor	Classifier	Detection			Evasion Rate (<i>MER</i>)			Feature Extractor	Classifier	Detection			Evasion Rate (<i>MER</i>)		
		P	R	F1	GAN & PSO	LSTM	Ours			P	R	F1	GAN & PSO	LSTM	Ours
CIC FLowMeter	KitNET	0.84	0.94	0.92	37.24%	50.69%	46.48%	CIC FLowMeter	KitNET	0.92	0.94	0.91	38.22%	44.39%	53.87%
	DT	0.79	0.91	0.82	35.88%	49.62%	60.98%		DT	0.74	0.79	0.76	49.98%	57.77%	64.12%
	IF	0.99	0.90	0.95	39.78%	42.93%	49.71%		IF	0.99	0.92	0.94	26.74%	38.64%	52.99%
	MLP	0.92	0.84	0.86	41.05%	53.87%	63.05%		MLP	0.77	0.72	0.74	52.07%	43.19%	73.21%
	SVM	0.99	0.92	0.95	88.79%	91.95%	64.49%		SVM	0.74	0.79	0.78	42.11%	45.54%	60.46%
	LR	0.84	0.91	0.89	24.72%	30.52%	42.08%		LR	0.73	0.78	0.72	35.87%	50.04%	47.48%

MALICIOUS FUNCTIONALITY

生成後もPort Scanの機能を維持



閉鎖ポートはRST/ACK、開放された**22番ポート**は**SYN/ACK**で応答。変換後のパケットも開放ポートを特定でき、Port Scanとして機能した。

PROACTIVE DEFENSE

攻撃トラフィック生成技術を防御の技術に

01 攻撃を生成

生成器を仮想的な攻撃者として使い、NIDSを回避する攻撃を作る。



04 再び攻撃

更新後のNIDSに対して攻撃を再生成し、再評価する。



02 弱点を発見

現在のNIDSへ入力し、見逃す攻撃とその条件を特定する。



03 防御を更新

見逃した生成データを追加し、NIDSを再学習する。



現実の攻撃者より先に
攻撃と防御を反復

現実の攻撃者より先に**攻撃と防御を反復**し、未知の回避攻撃へ備える。
つまり、「**いたちごっこ**」を先にやる。

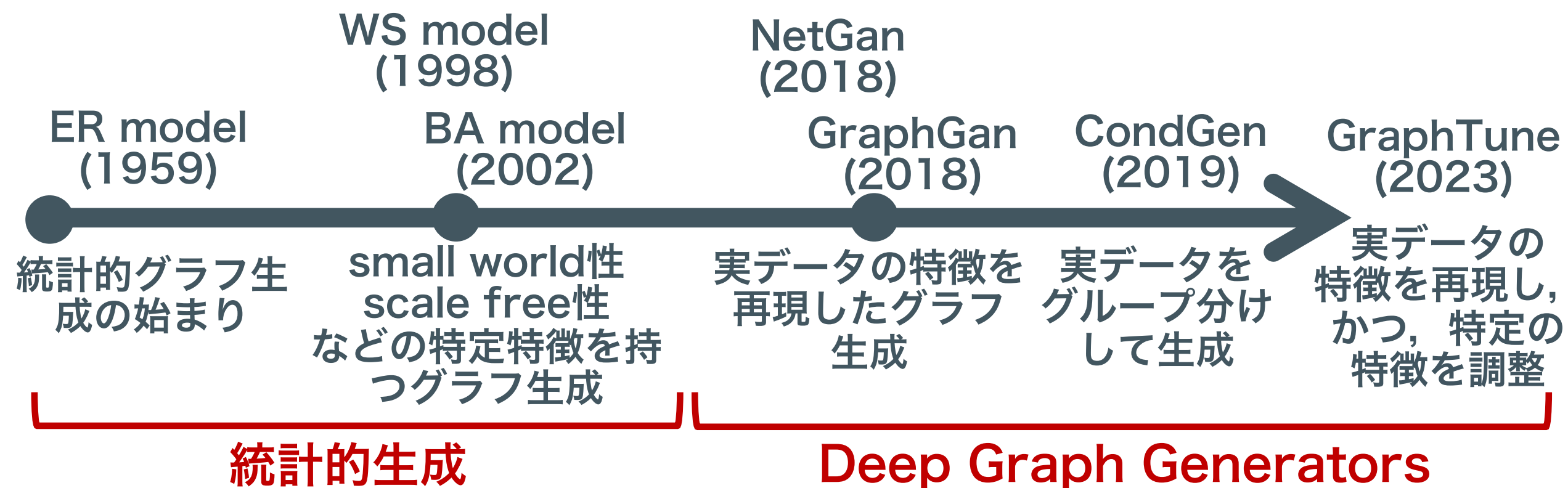
PART III

グラフ生成モデルの発展

HISTORY

グラフ生成技術と特徴量の発展

グラフ生成技術は**1959年のERモデル**に始まり、現代に至る。



特徴を定量化する指標

- **距離**：平均最短経路長、直径
- **構造**：クラスタリング係数、モジュラリティ
- **接続**：平均次数、次数分布のべき指数、エッジ密度
- **重要度**：固有値、次数中心性、媒介中心性、PageRank

STATISTICAL MODEL

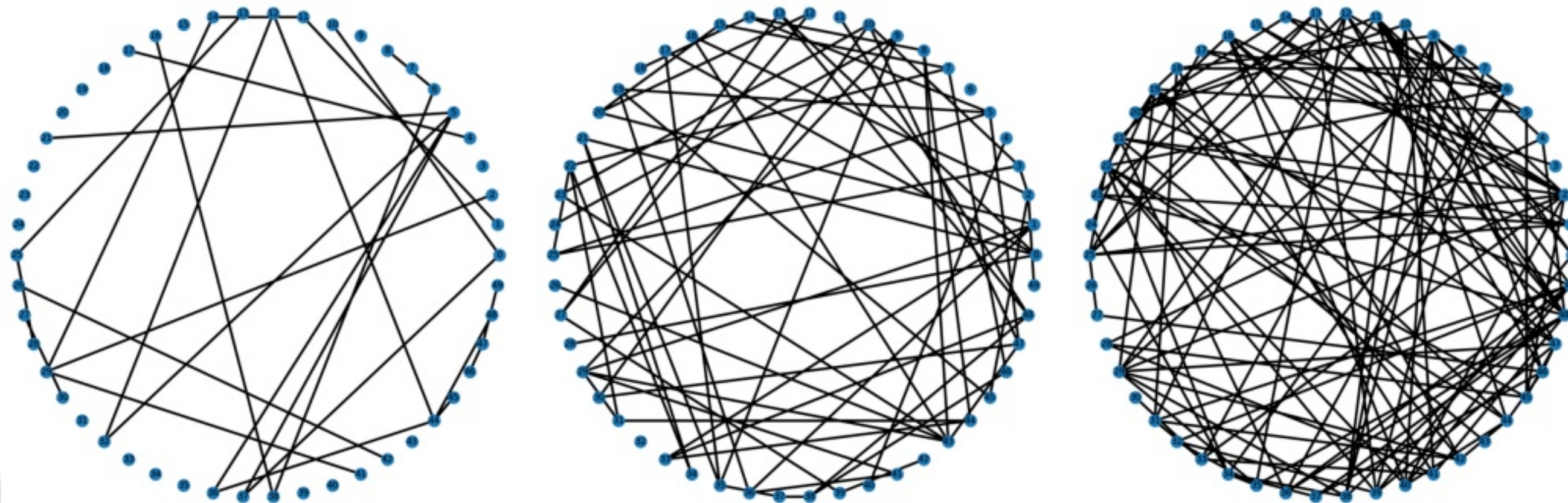
統計的な生成モデル：ERモデル

Erdős–Rényiモデルは、最も原始的なグラフ生成モデル。

入力パラメータ：ノード数 n 、エッジ密度 p

1. n 個のノードを作成する。
2. すべてのノード対を確率 p で接続する。

エッジ密度を変えた生成例



STATISTICAL MODEL

統計的な生成モデル：WSモデル

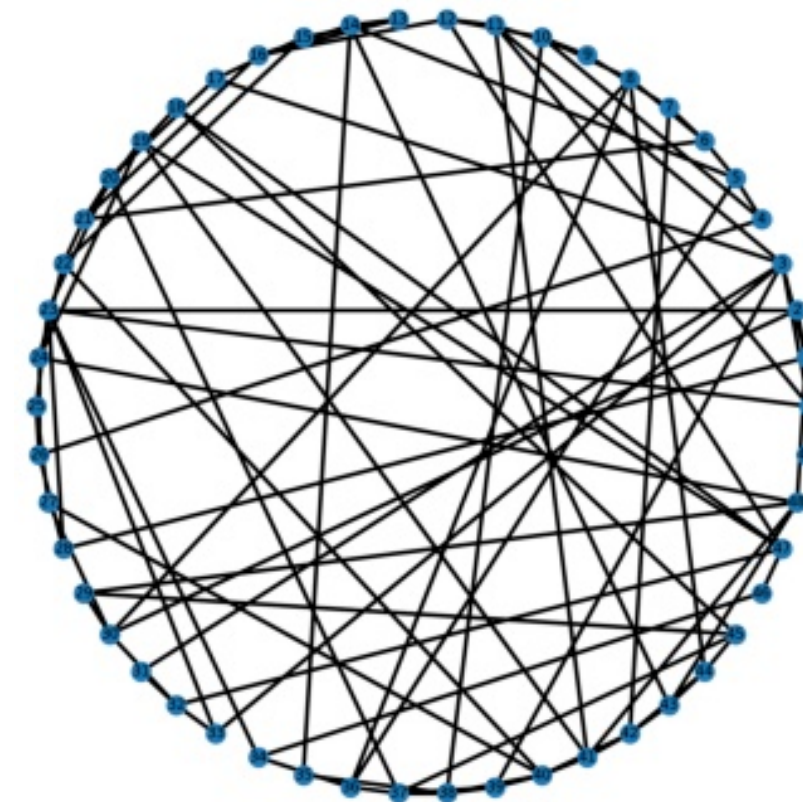
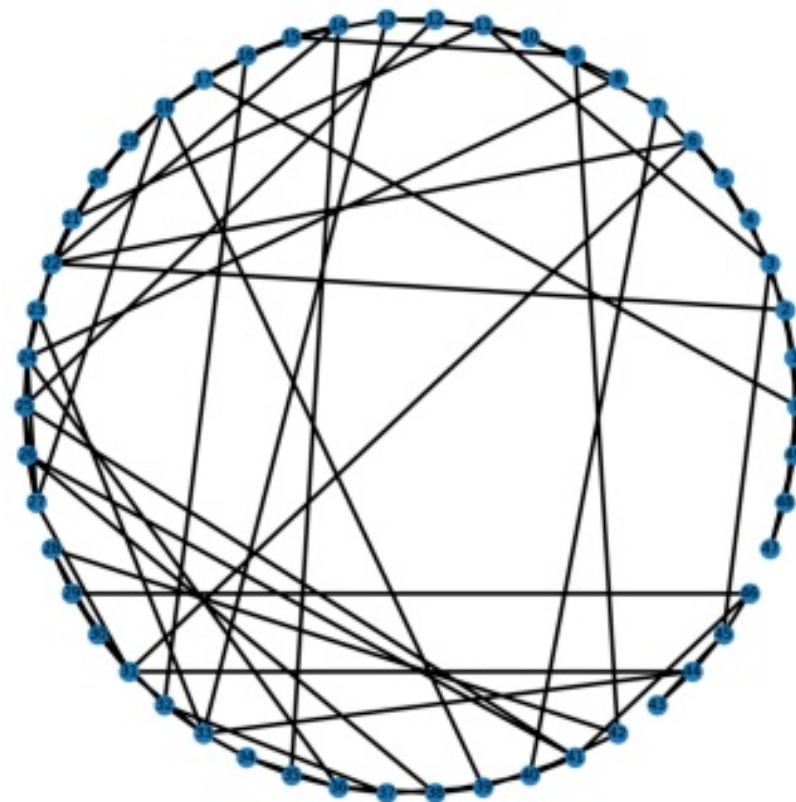
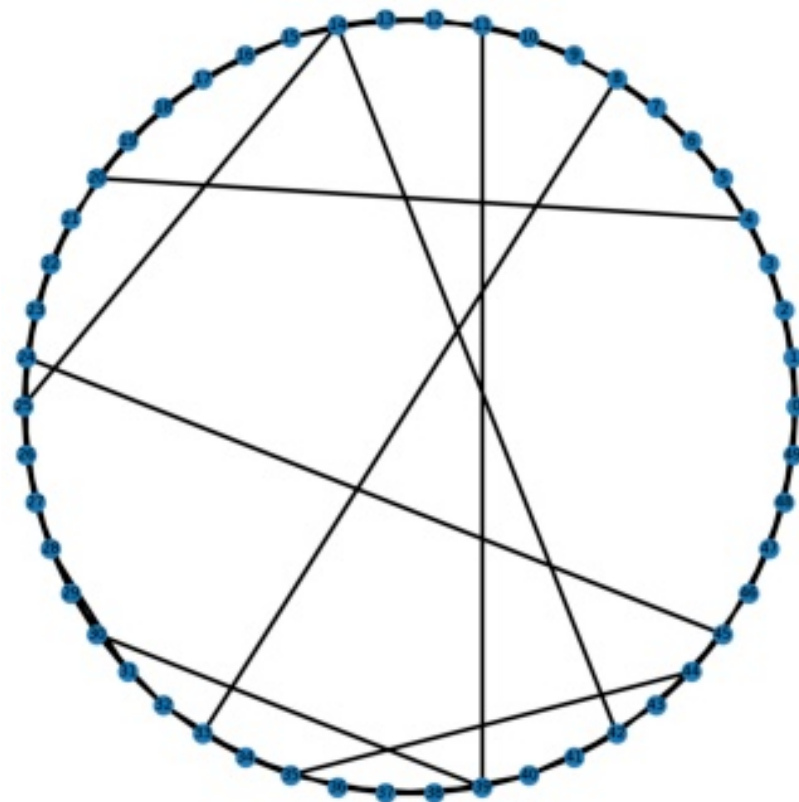
Watts–Strogatzモデルは、スモールワールド性を再現する。

スモールワールド性：ノード数 n に対し、平均経路長は $L = O(\log n)$ で増加する。

エッジ張替え確率を変えた生成例

入力パラメータ：ノード数 n 、平均次数 $2K$ 、張替え確率 p

1. 平均次数が $2K$ のリング格子を作る。
2. 各エッジを確率 p で張り替える。



STATISTICAL MODEL

統計的な生成モデル：BAモデル

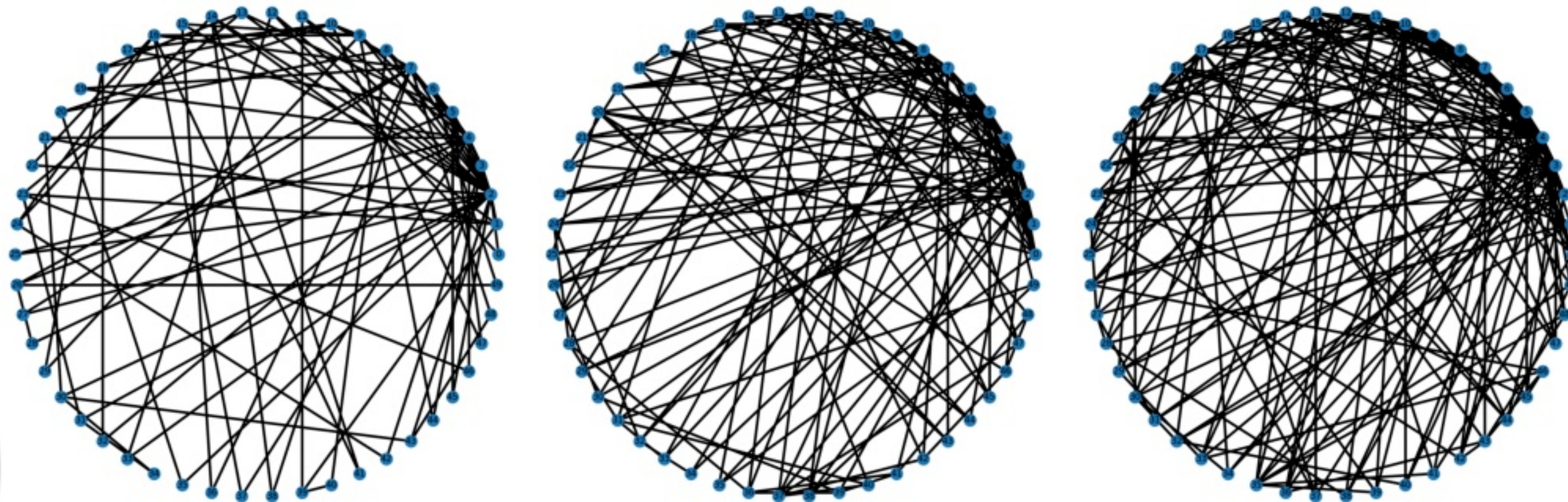
Barabási–Albertモデルは、スケールフリー性を再現する。

スケールフリー性：次数分布が $f(k) \propto k^{-\gamma}$ ($2 \leq \gamma \leq 3$) となる。

追加エッジ数を変えた生成例

入力パラメータ：追加エッジ数 m 、ノード数 n

1. m 個のノードから完全グラフを作る。
2. m 本のエッジを持つノードを追加する。
3. 接続先は既存ノードの次数に比例して選ぶ。

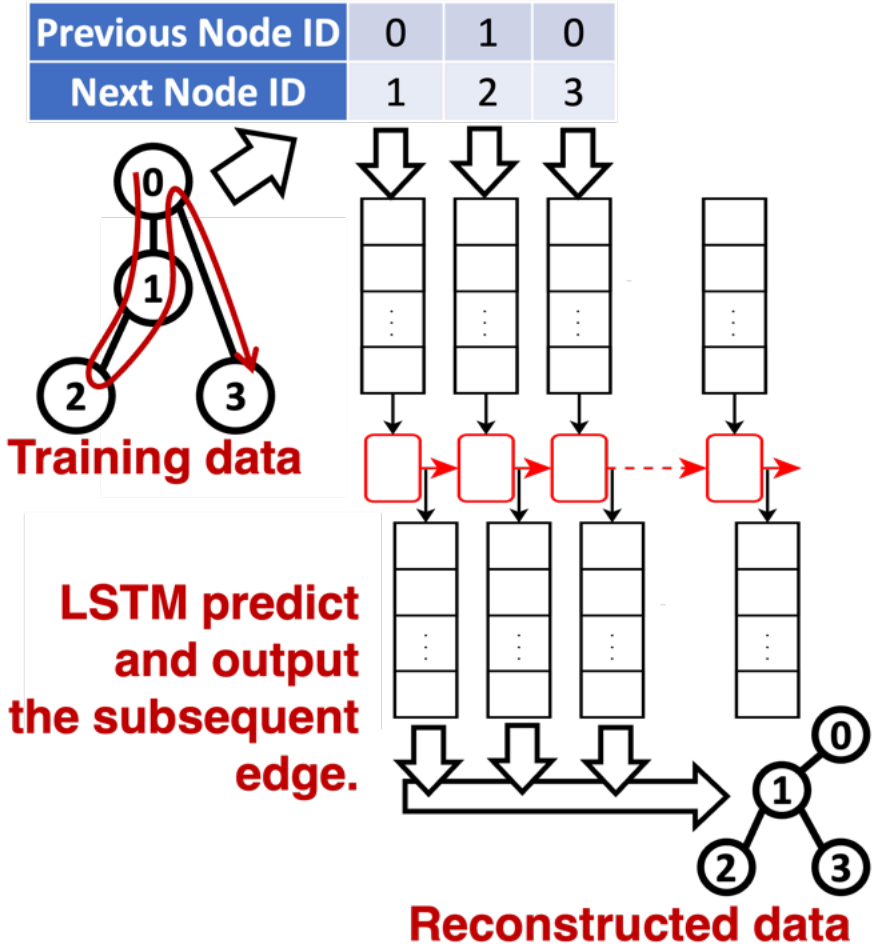


DEEP GRAPH GENERATOR

GraphGen: 実グラフを系列として学習

統計的な生成規則の代わりに、**実グラフの構造を深層学習で模倣する。**

- 1. DFSでグラフをエッジ系列へ変換
- 2. LSTMへ系列を入力し、次のエッジを予測
- 3. 学習済みモデルから系列を再帰的に生成
- 4. 生成系列をグラフへ逆変換

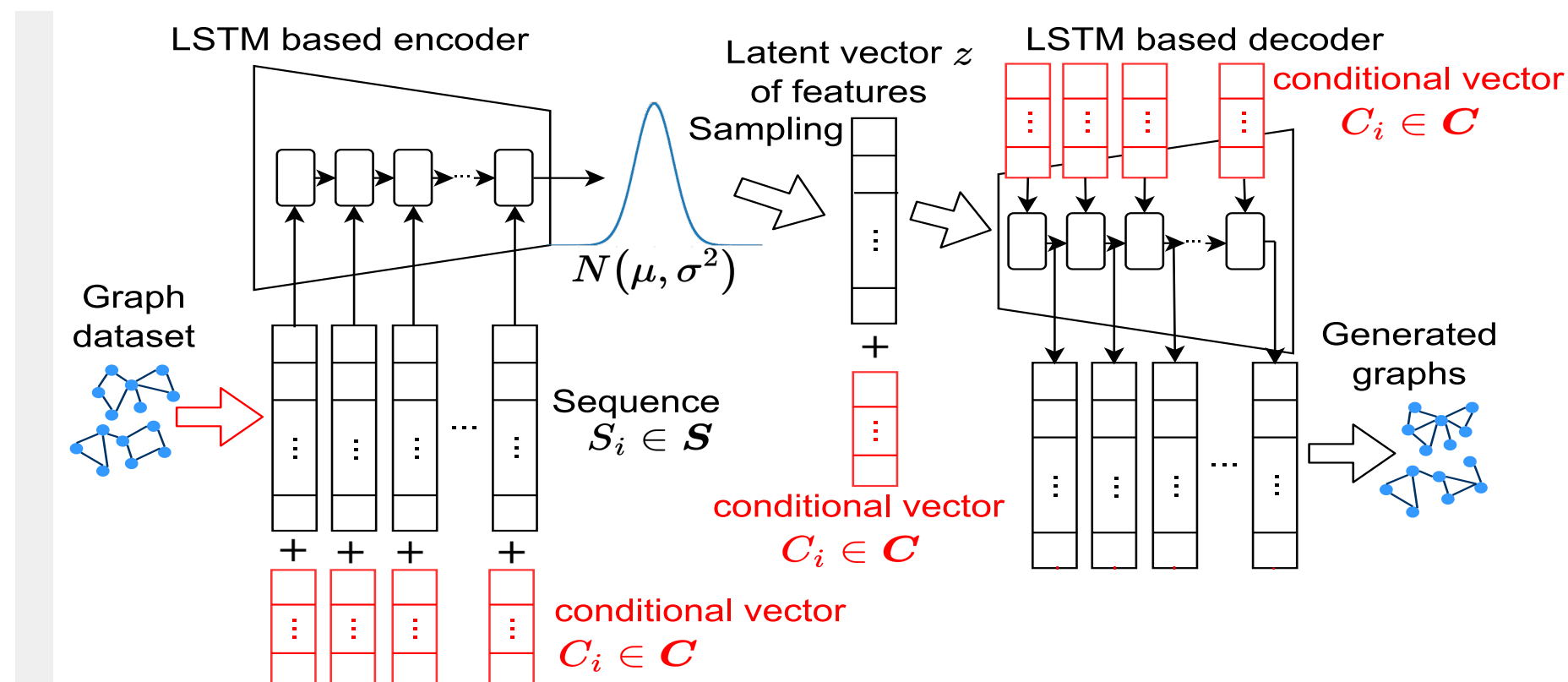


CONDITIONAL GENERATION

GraphTune: 特徴量を指定して生成

CVAEと**LSTM**を組み合わせ、生成グラフの特徴量を条件として指定する。

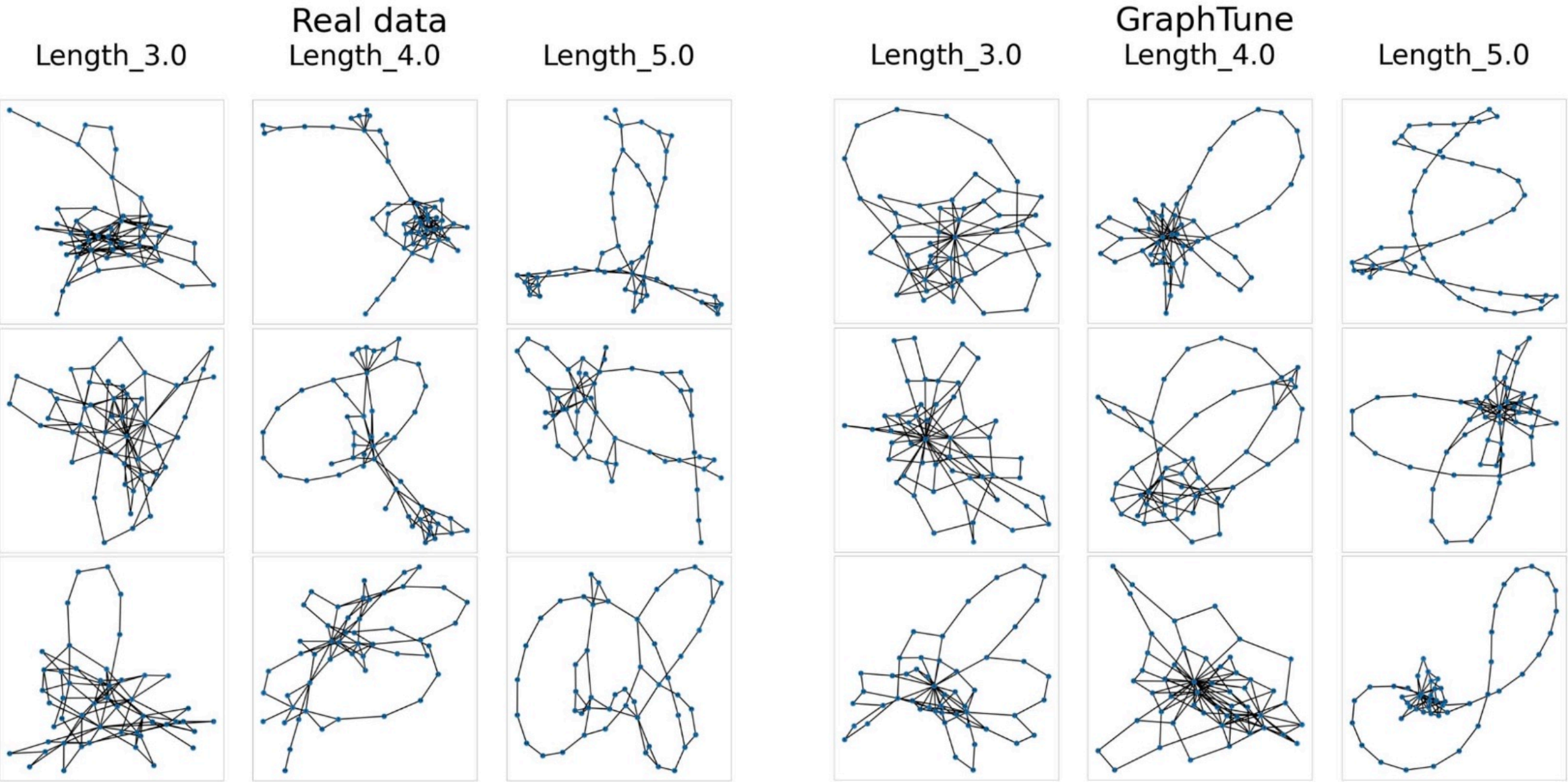
グラフ系列を潜在空間へ圧縮し、条件ベクトルを加えて復号することで**連続的な調整**を実現する。



K. Watabe, S. Nakazawa, Y. Sato, S. Tsugawa, and K. Nakagawa, “GraphTune: A Learning-Based Graph Generative Model With Tunable Structural Features,” IEEE Transactions on Network Science and Engineering, vol. 10, no. 4, 2023.

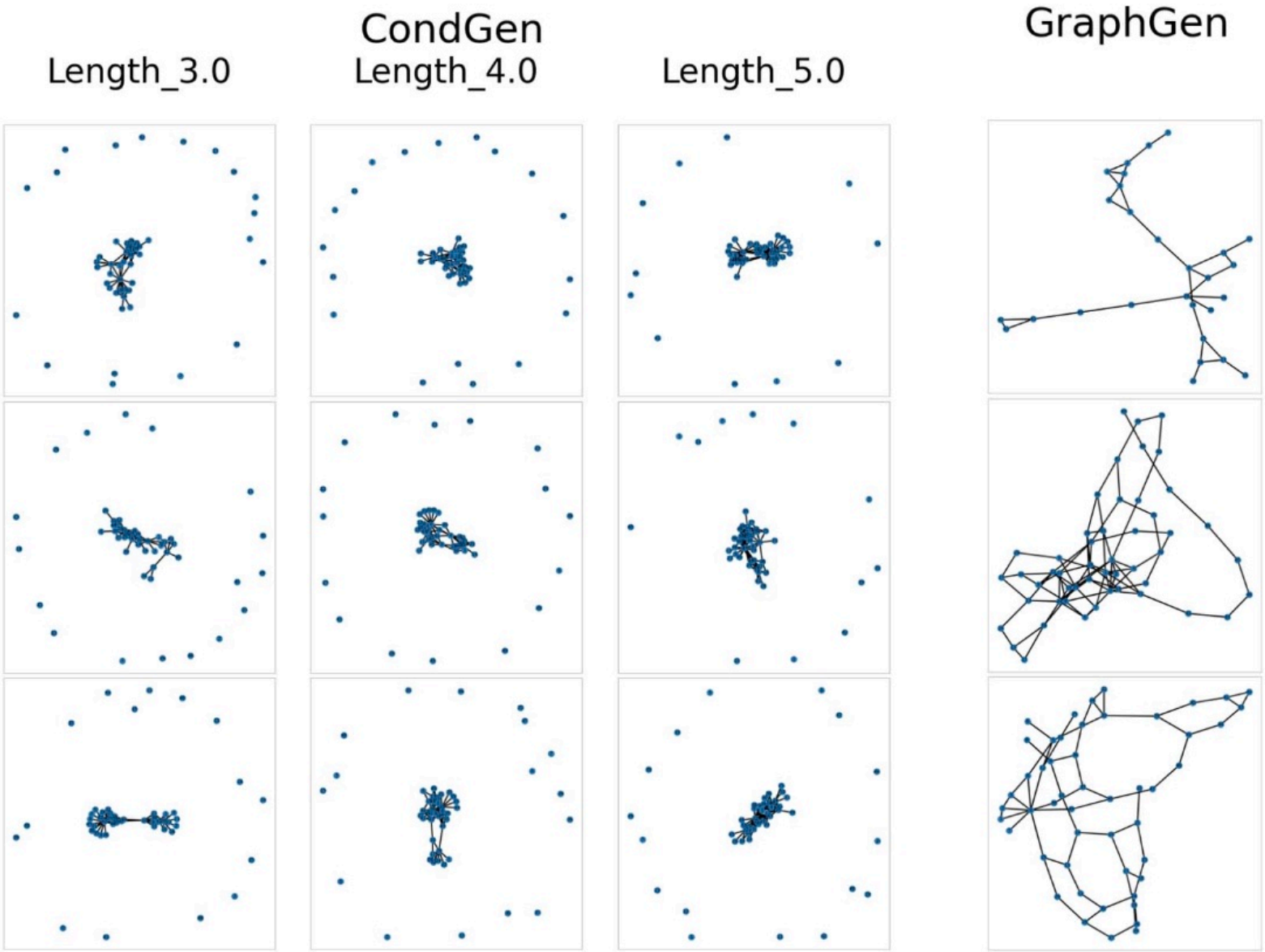
GENERATED GRAPH EXAMPLES

Real vs GraphTune



BASELINE MODELS

CondGen and GraphGen

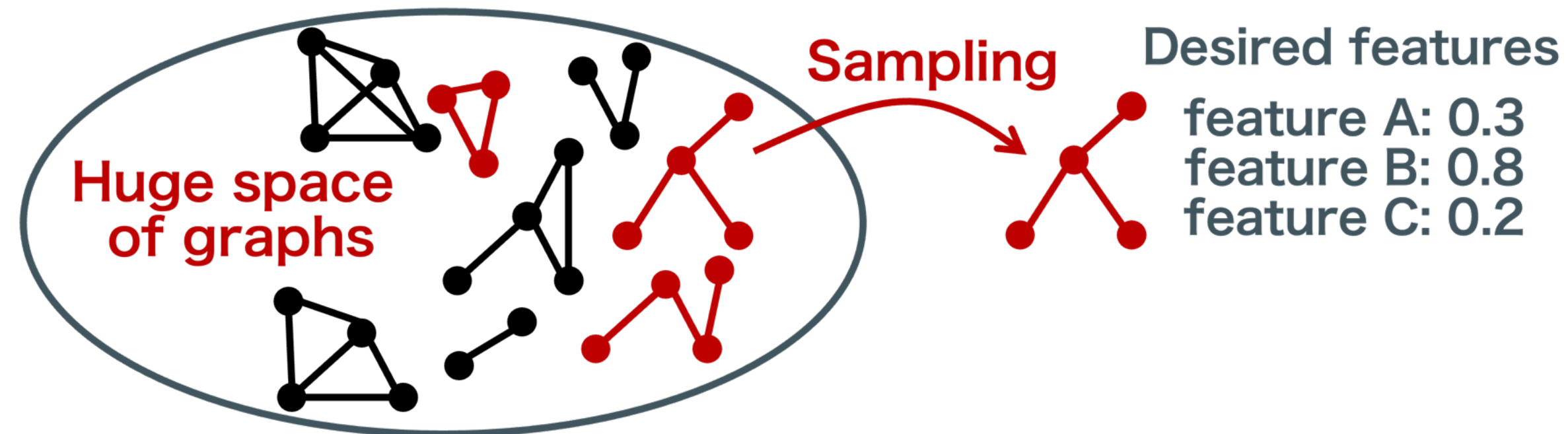


OBJECTIVE

グラフ生成技術の目的

広大なグラフ空間から、**所望の特徴を持つグラフ**をどうサンプリングするか。

WSモデル	スモールワールド性を再現する（平均最短経路長）。	BAモデル	スケールフリー性を再現する（次数分布のべき指数）。
GraphGen	学習データのグラフと同じ特徴を再現する。	GraphTune	学習データの他の特徴を再現しながら、指定した特徴を調整する。



PART IV

生成データの品質評価

QUALITY ASSURANCE

生成データの品質をどう保証するか

目標は、**人手による個別検証を挟まず**、生成データをシステムへ投入すること。

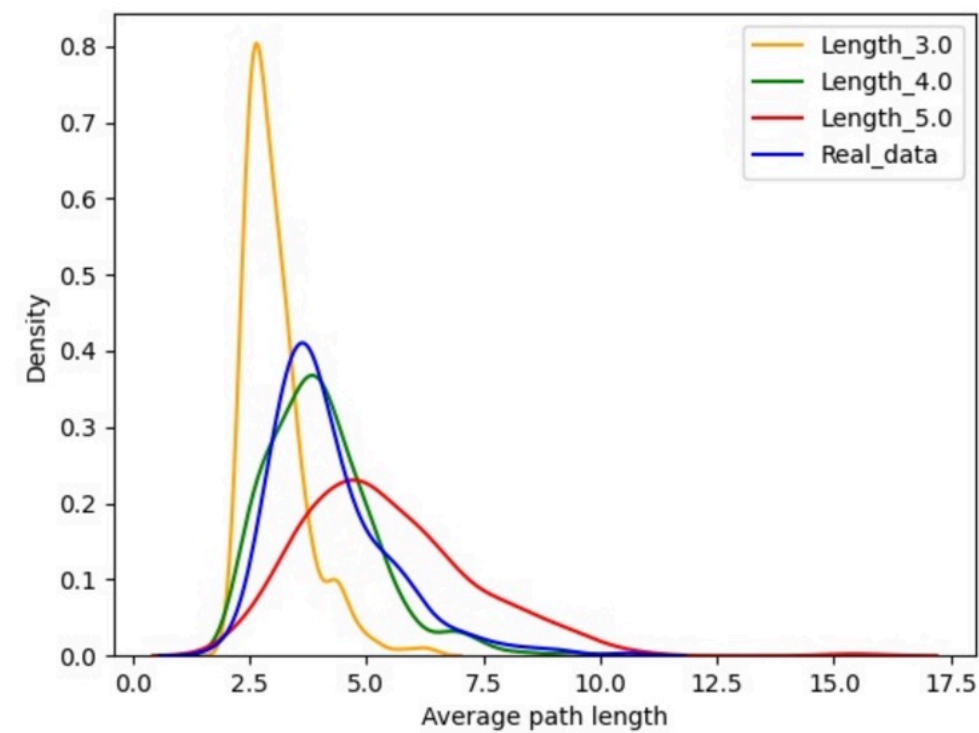
01 制御精度 指定特徴をばらつきなく再現できるか 生成結果のばらつきを抑え、指定した特徴値へ高精度に一致させる必要がある。	02 多様性 同じ条件でも異なるデータを生成できるか 一部のパターンへ偏らず、実データが持つ多様な構造を生成する必要がある。	03 類似性 何を測れば「似ている」といえるか 分布、相関、構造特徴などの類似性をどう測り、実データ相当の品質をどう保証するか。
---	---	---

CONTROL ACCURACY

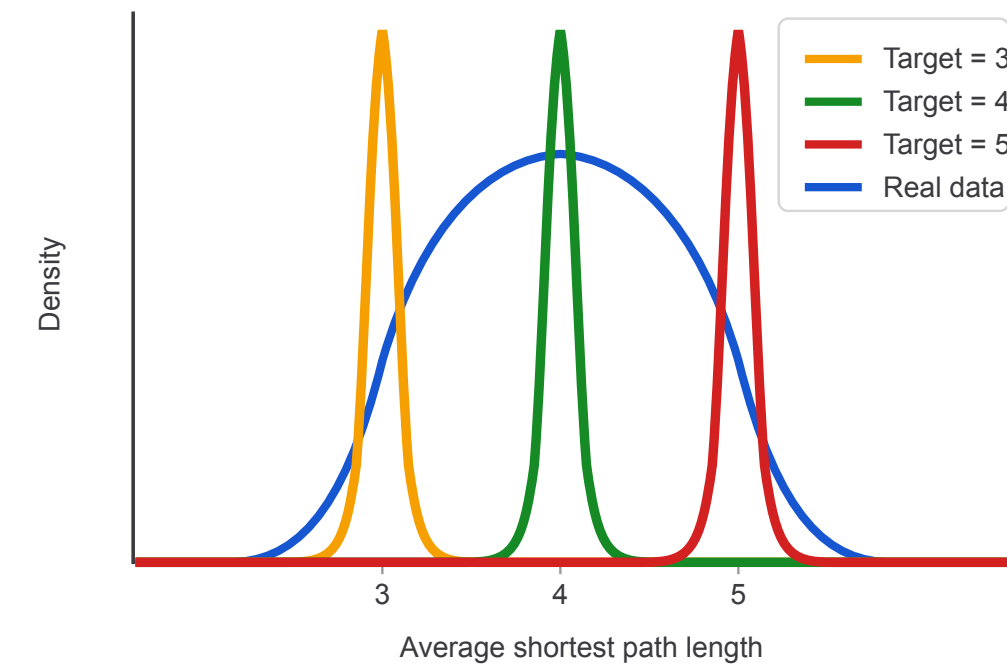
制御精度の評価例

指定値に対する**ピーク位置のずれ**と**分布の広がり**を比較する。

生成結果の分布



理想的な分布

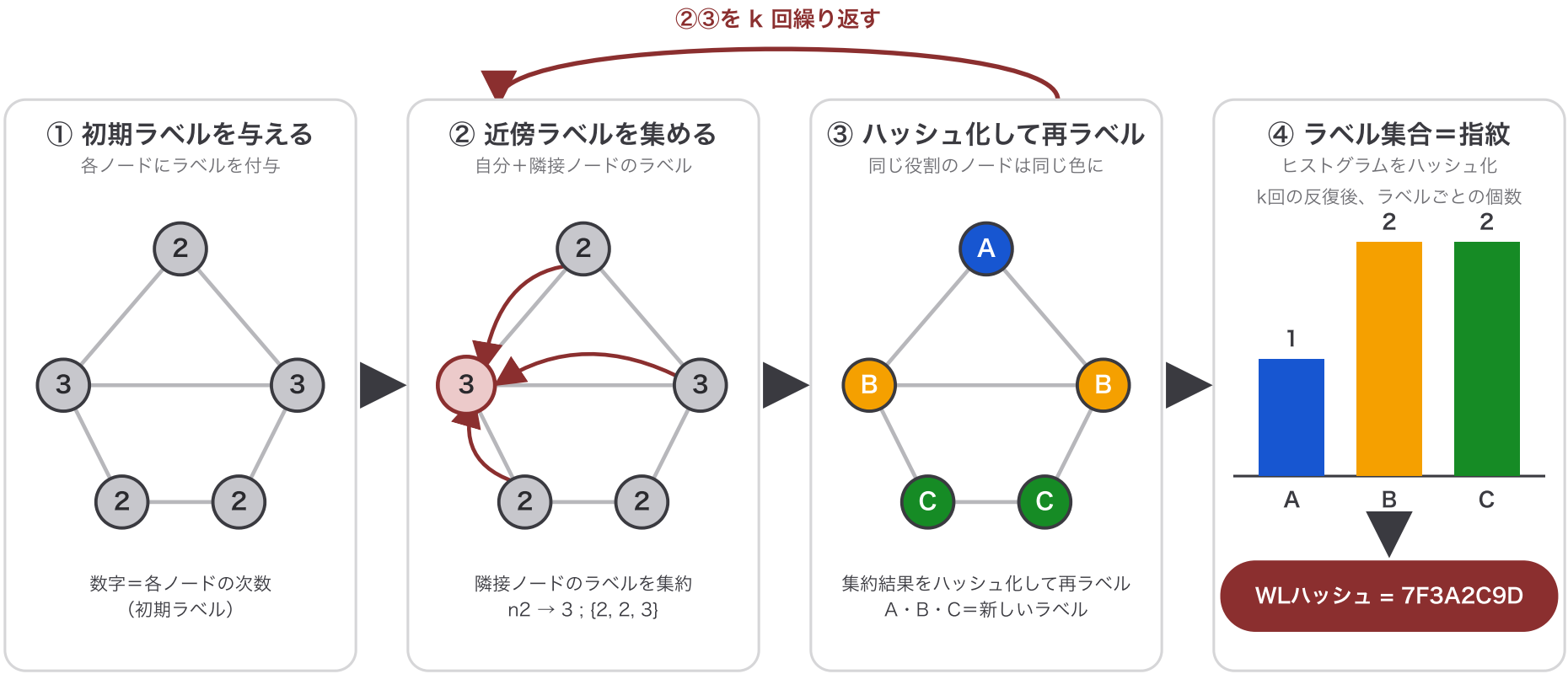


理想は、**重なりの少ない急峻なピーク**が現れる状態。

DIVERSITY

多様性はWeisfeiler-Lehmanテストで調べられる

生成グラフ同士が構造的に重複していないかは、**グラフをハッシュ化して比較**することで確かめられる。



指紋（WLハッシュ）が生成グラフ間で**集中して一致**すれば、多様性の低下が疑われる。

MULTIDIMENSIONAL EVALUATION

特徴量は多面的に評価する

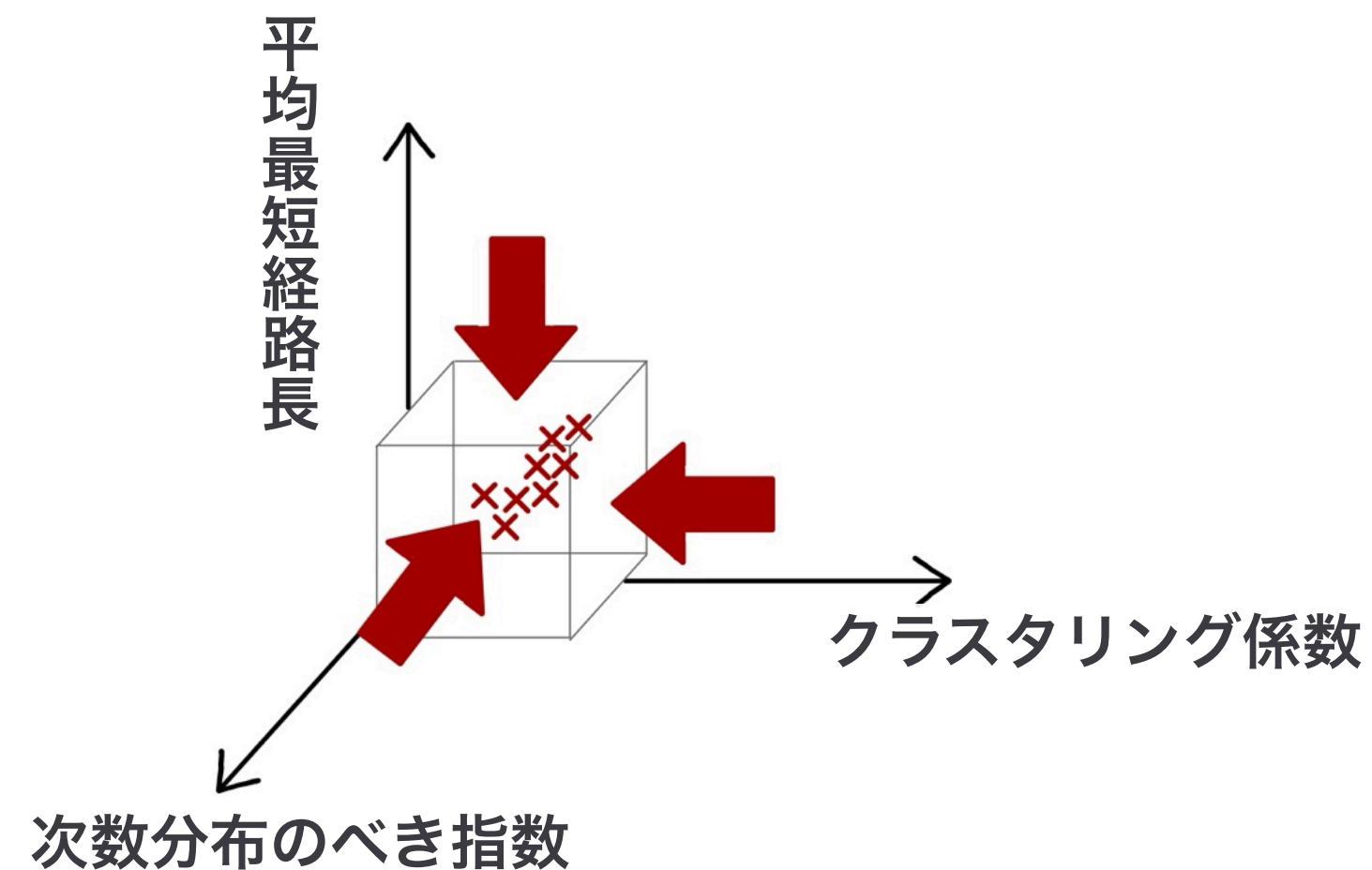
一つの特徴量が一致しても、グラフ全体が似ているとは限らない。

距離	平均最短経路長
----	---------

局所構造	クラスタリング係数
------	-----------

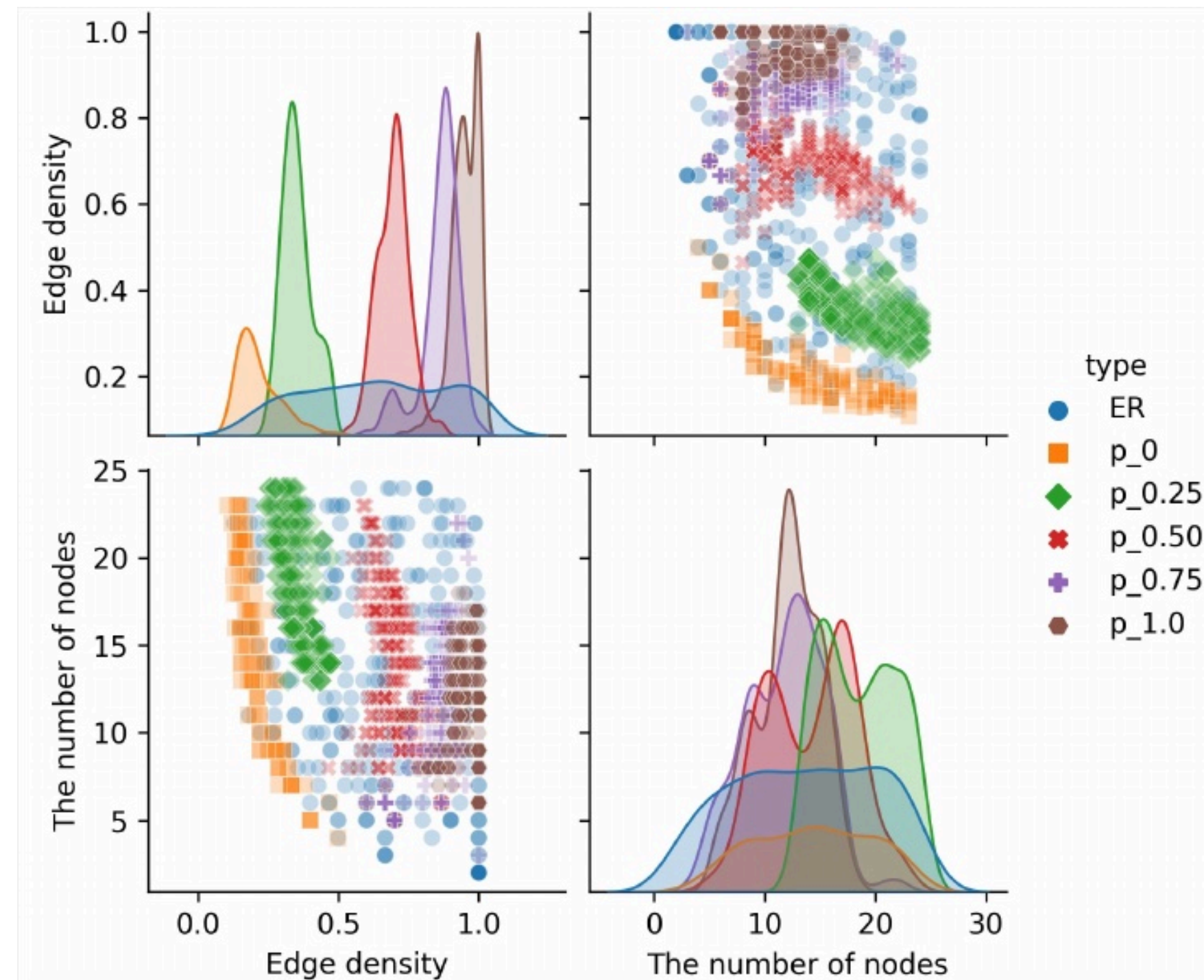
接続構造	次数分布のべき指数
------	-----------

複数の特徴量を組み合わせ、異なる方向から同時に確認する。



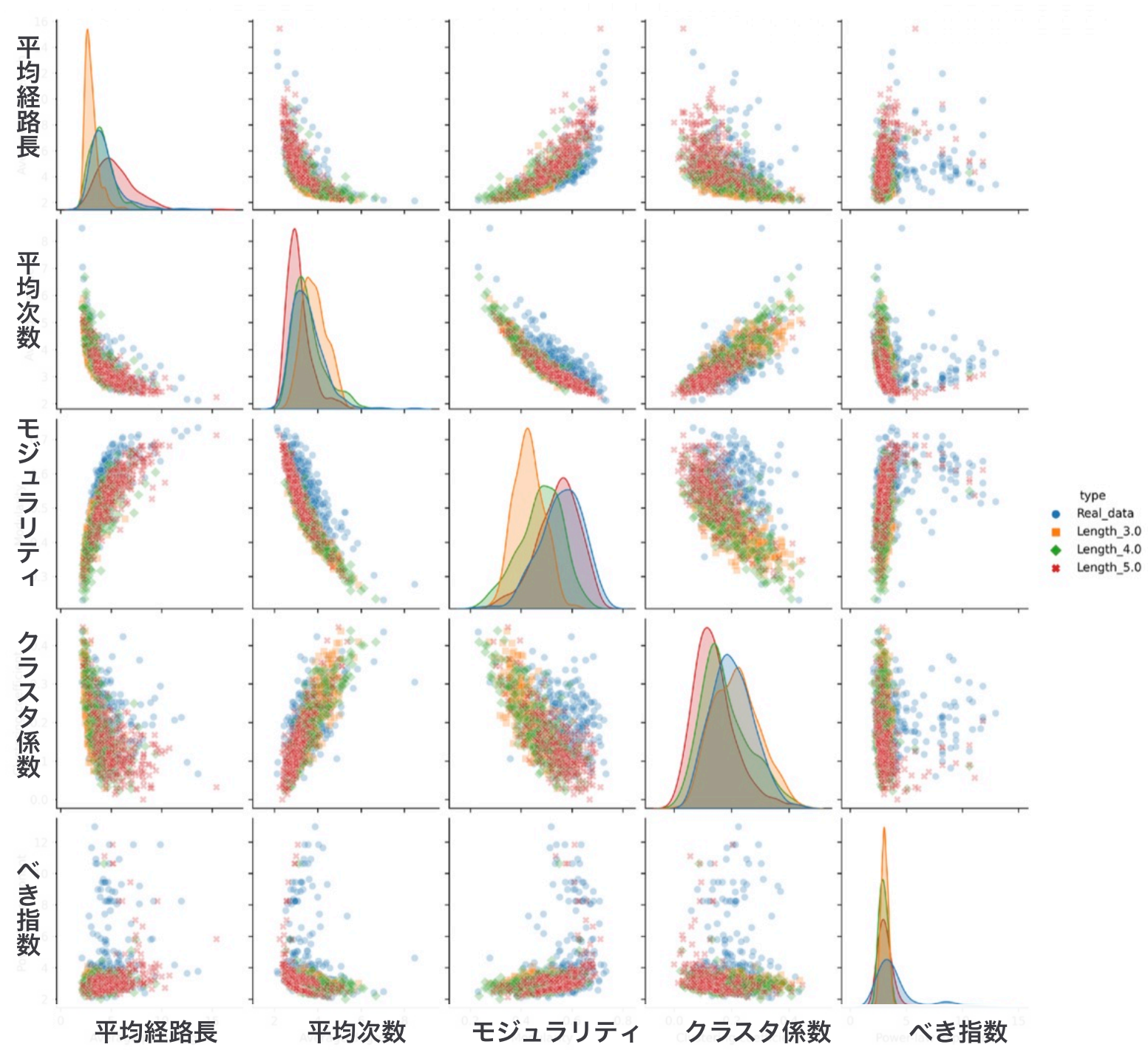
GENERATED FEATURE SPACE

ERグラフと生成グラフの特徴空間



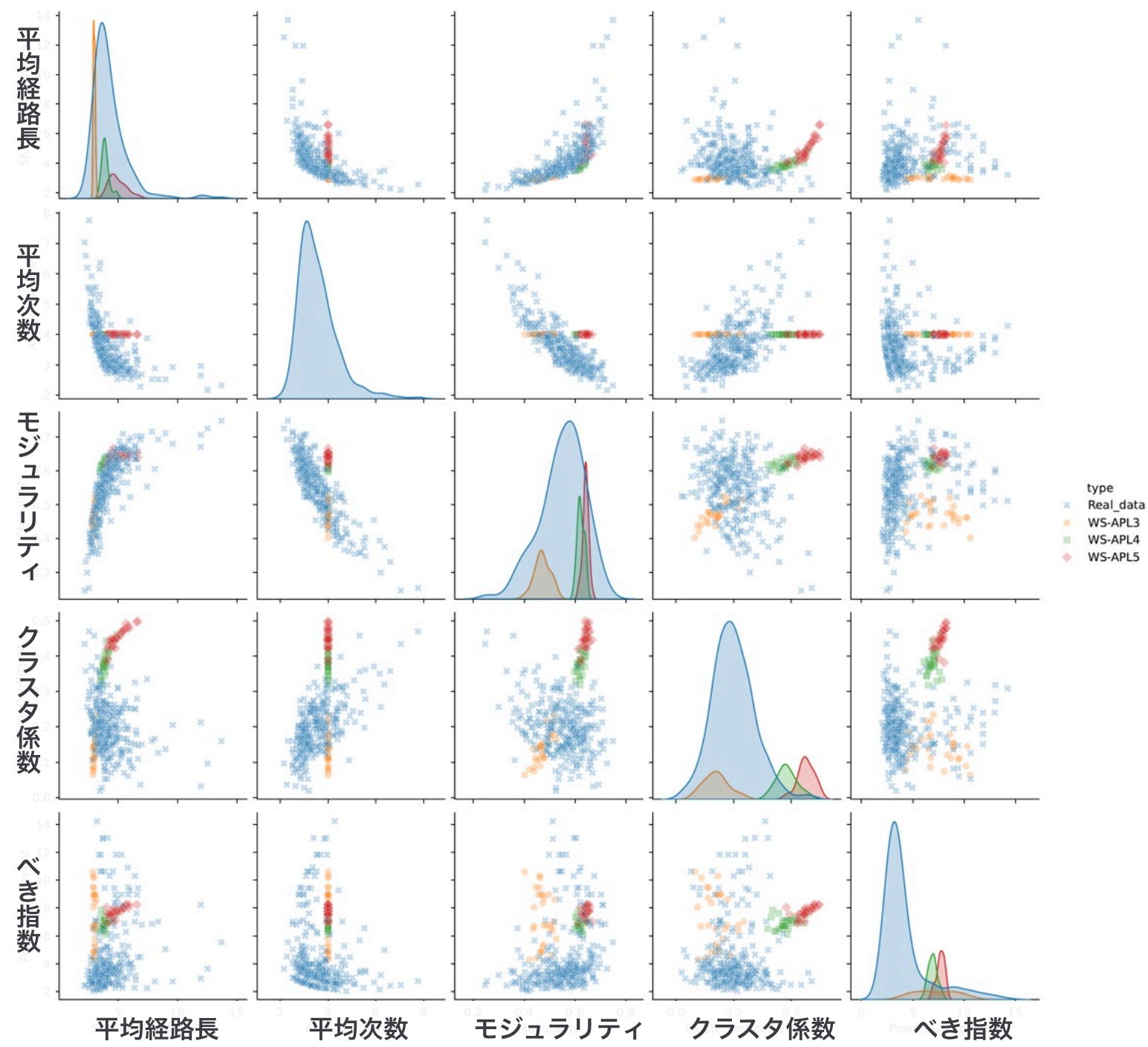
FEATURE DISTRIBUTIONS

GraphTuneの分布



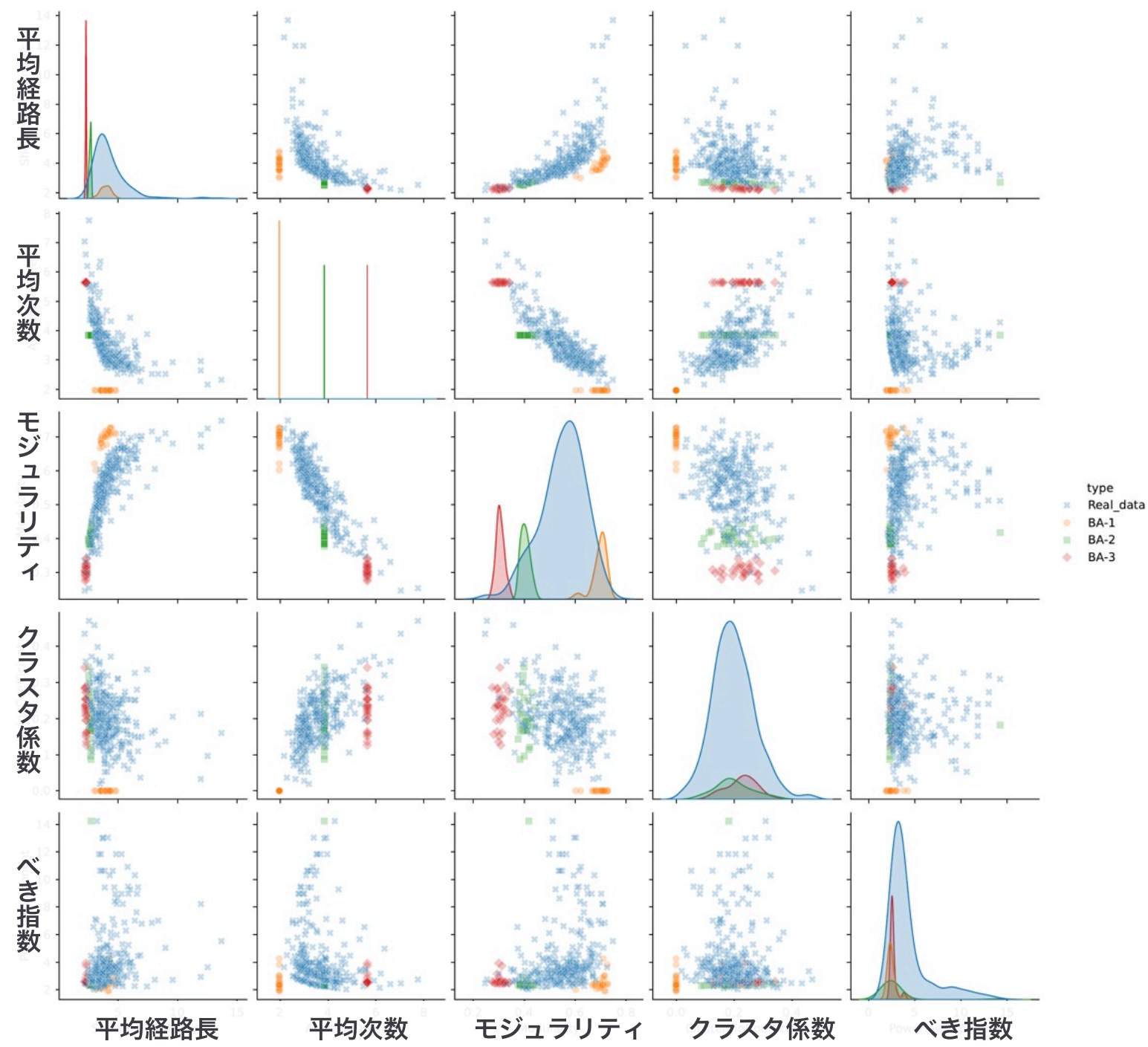
WS MODEL

WSモデルと実データの特徴量分布



BA MODEL

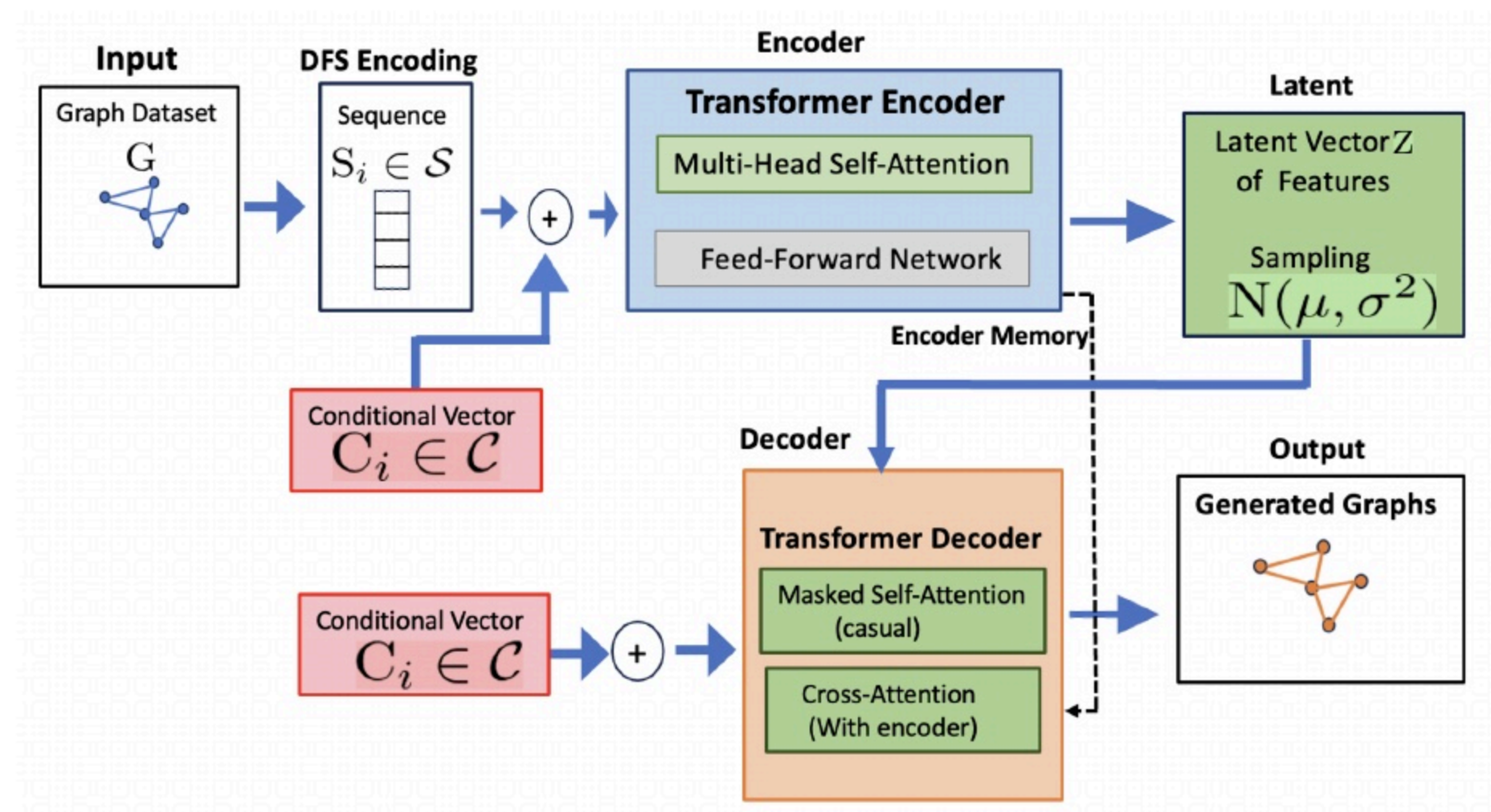
BAモデルと実データの特徴量分布



T-GRAPHTUNE

NNのリッチ化による生成精度の向上

従来のLSTMを**Transformer**へ置き換え、自己注意により系列全体の依存関係を捉える。

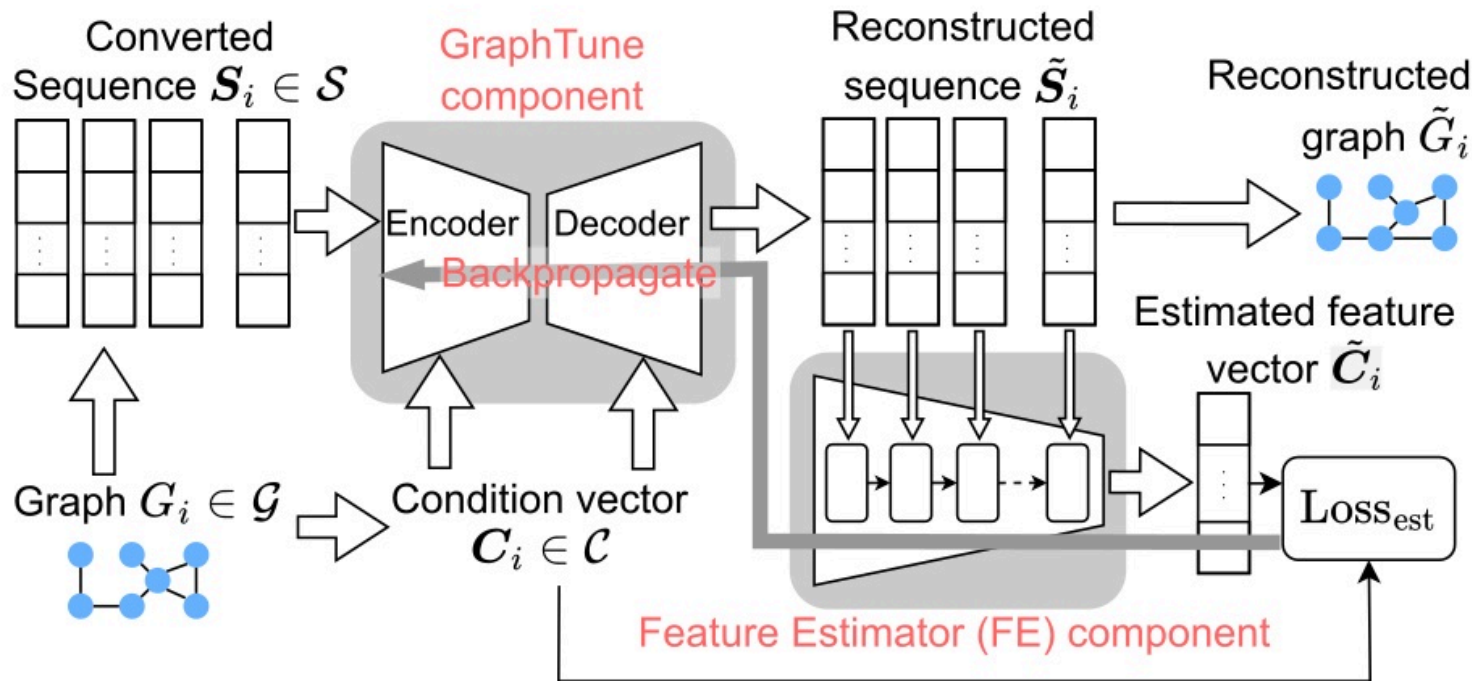


FEATURE FEEDBACK

F-GraphTune：特徴量をフィードバック

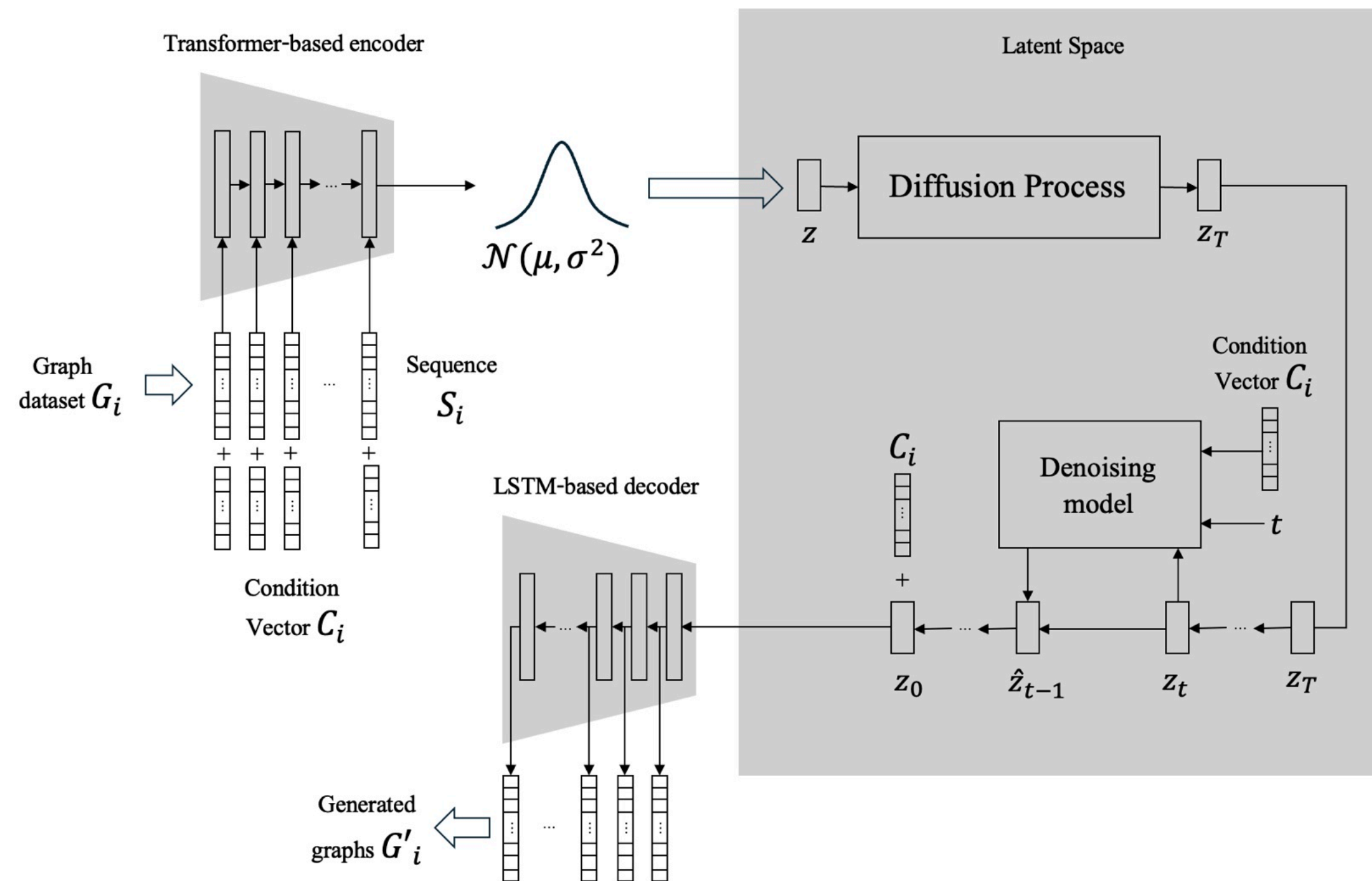
再構築した系列から、**Feature estimator**がグラフの特徴量を推定する。

推定値と指定値の差をGraphTuneへ戻し、両モデルを交互に学習して制御精度を高める。



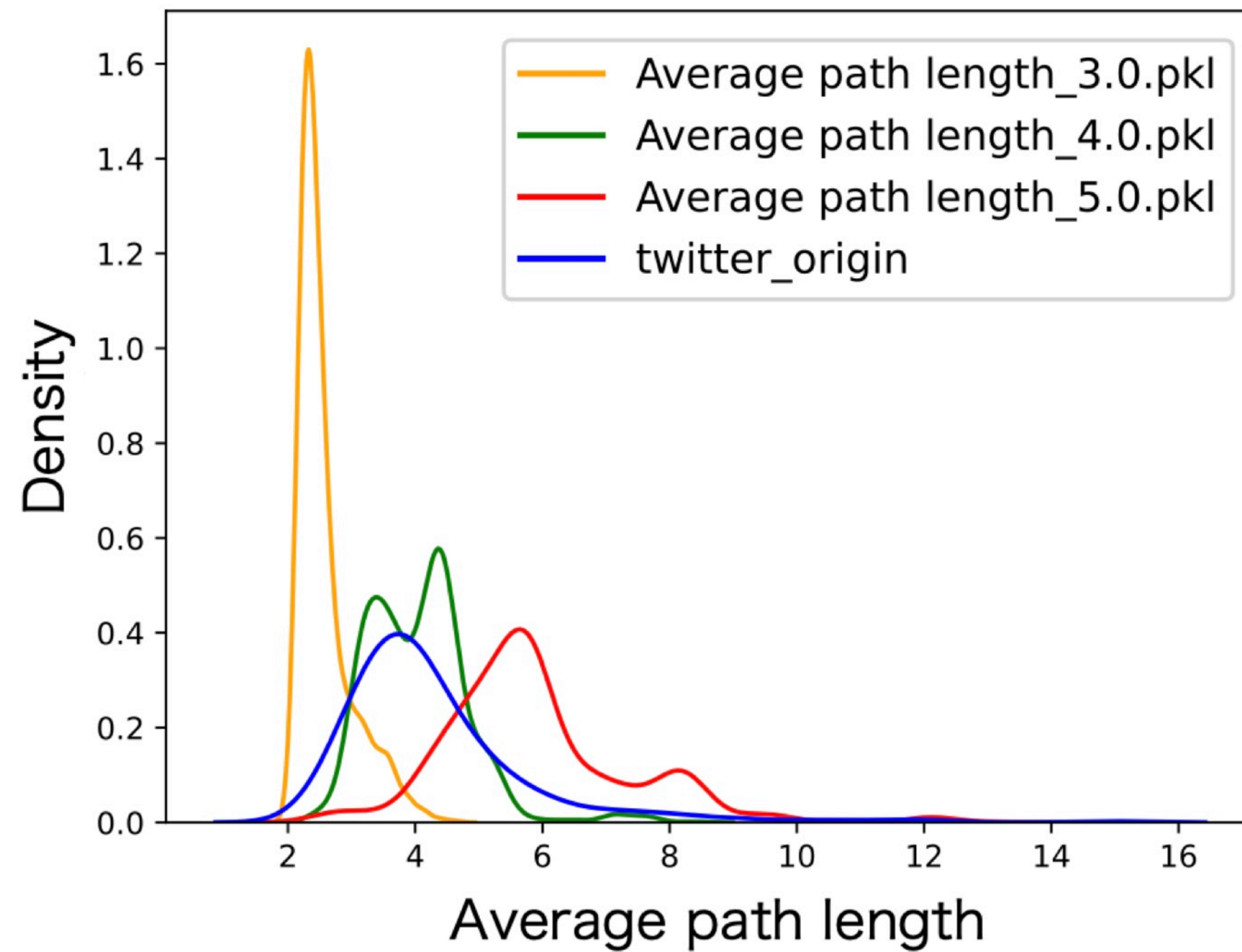
DIFFUSION-BASED GENERATION

潜在空間に拡散過程を導入する



CONTROLLED DISTRIBUTIONS

指定した平均最短経路長ごとの生成分布



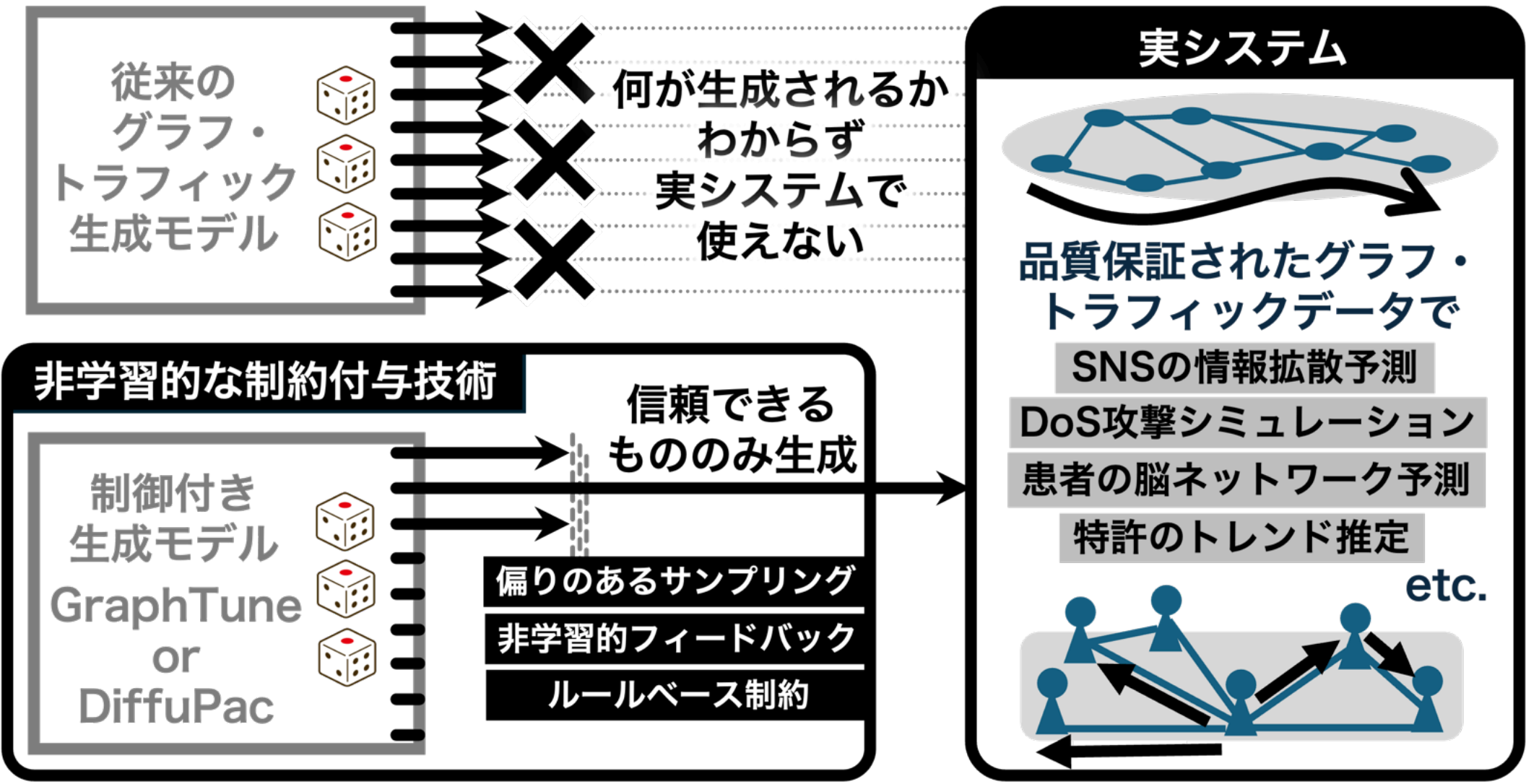
PART V

モデルの複雑性の弊害

BLACK-BOX GENERATION

高性能でも生成はブラックボックスでばらつく

条件を指定しても、乱数によって個々の出力は変わる。**何が生成されるかを事前に保証できない。**



CONTROL AND GUARANTEE

制御から品質保証へ

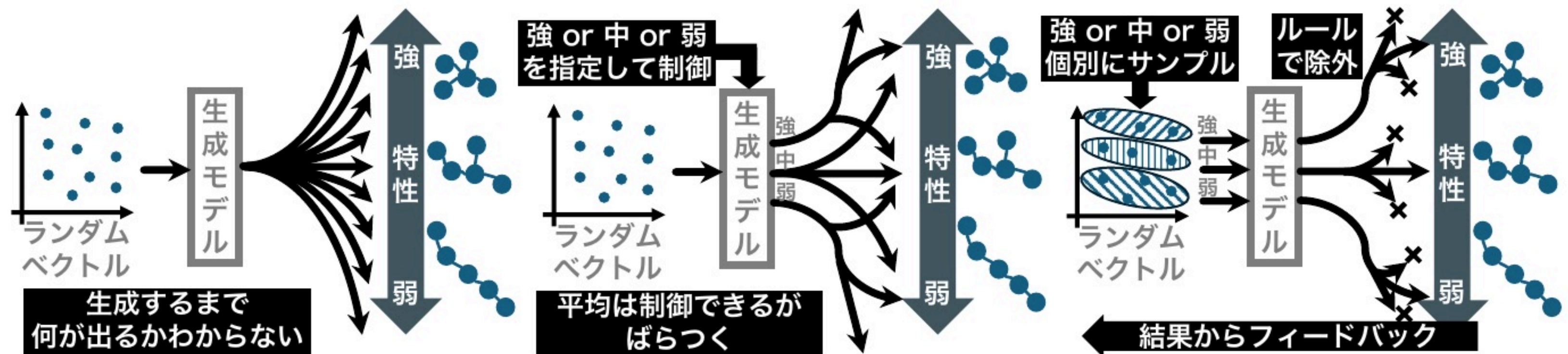


図 2: (左) 無制御 (従来法), (中央)GraphTune/DiffuPac の制御, (右) 制約を掛けた制御

TRAINING-FREE CONSTRAINTS

生成モデルの外側から品質を制約する

サンプリング空間

条件を満たす潜在領域へ
探索範囲を限定する

非学習フィードバック

生成結果を測定し、次の
サンプル位置へ反映する

ルール制約

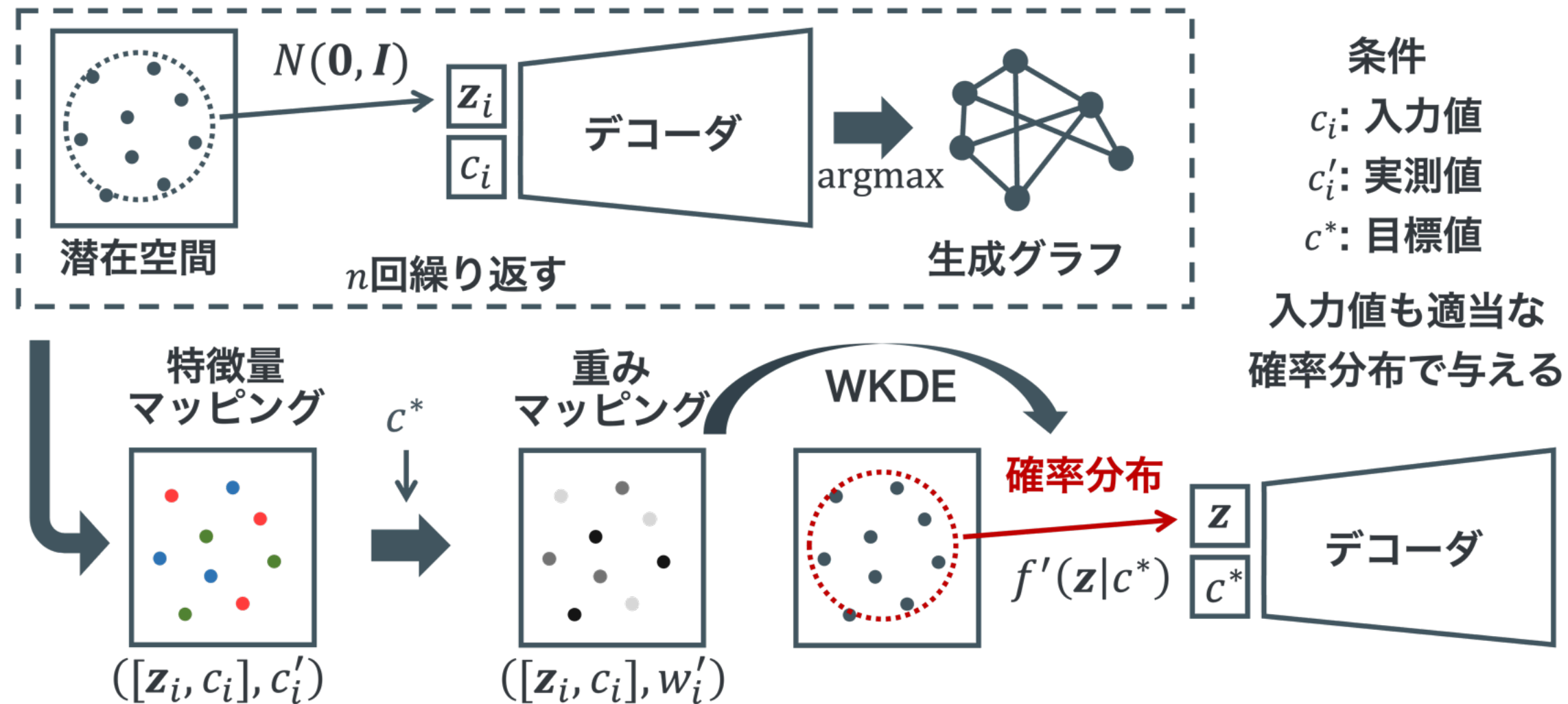
連結性やプロトコル条件
に反する結果を除外する

追加学習を避け、ブラックボックスな生成器にも同じ制約層を適用する。

TRAINING-FREE FEEDBACK

WKDEでランダムベクトルの分布にバイアスを掛ける

適当な分布から z と c をサンプルして生成し、**実測値との対応をWKDEで確率分布に変換する。**



目標条件 c^* に近い z ほど選ばれやすい分布を、生成器を再学習せずに得る。

SAMPLING RESULT

WKDEによるヒット率の改善

目標APLに対し、許容誤差 ϵ 以内に収まったサンプル数を1000サンプル中で比較する。

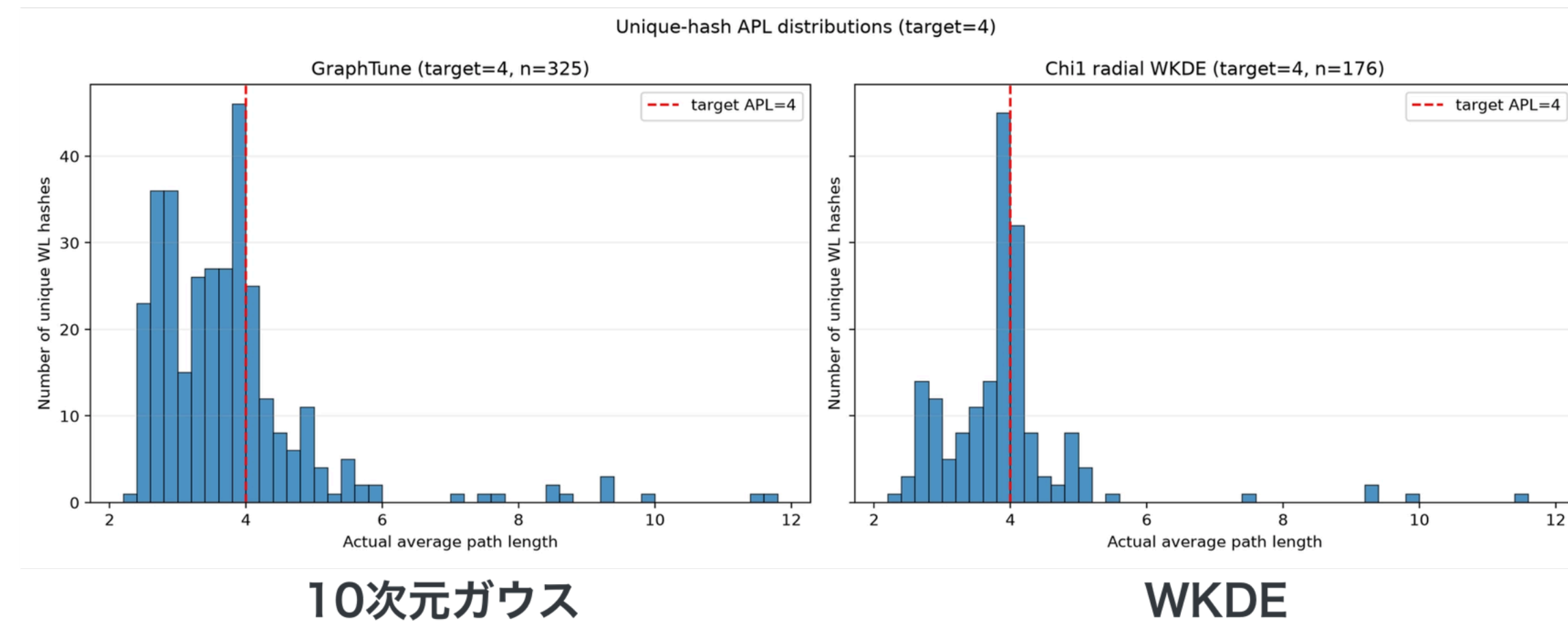
目標APL	サンプリング手法	$\epsilon=0.1$	$\epsilon=0.2$	$\epsilon=0.3$	ALL
3.0	10次元ガウス	14	61	76	1000
3.0	WKDE	95	235	308	1000
4.0	10次元ガウス	106	322	340	1000
4.0	WKDE	433	675	708	1000
5.0	10次元ガウス	26	68	134	1000
5.0	WKDE	236	429	482	1000

目標値や ϵ によらず、WKDEは10次元ガウスより多くのサンプルが目標条件を満たす。

CONTROL VS. DIVERSITY

WKDEは精度と引き換えに多様性がやや下がる

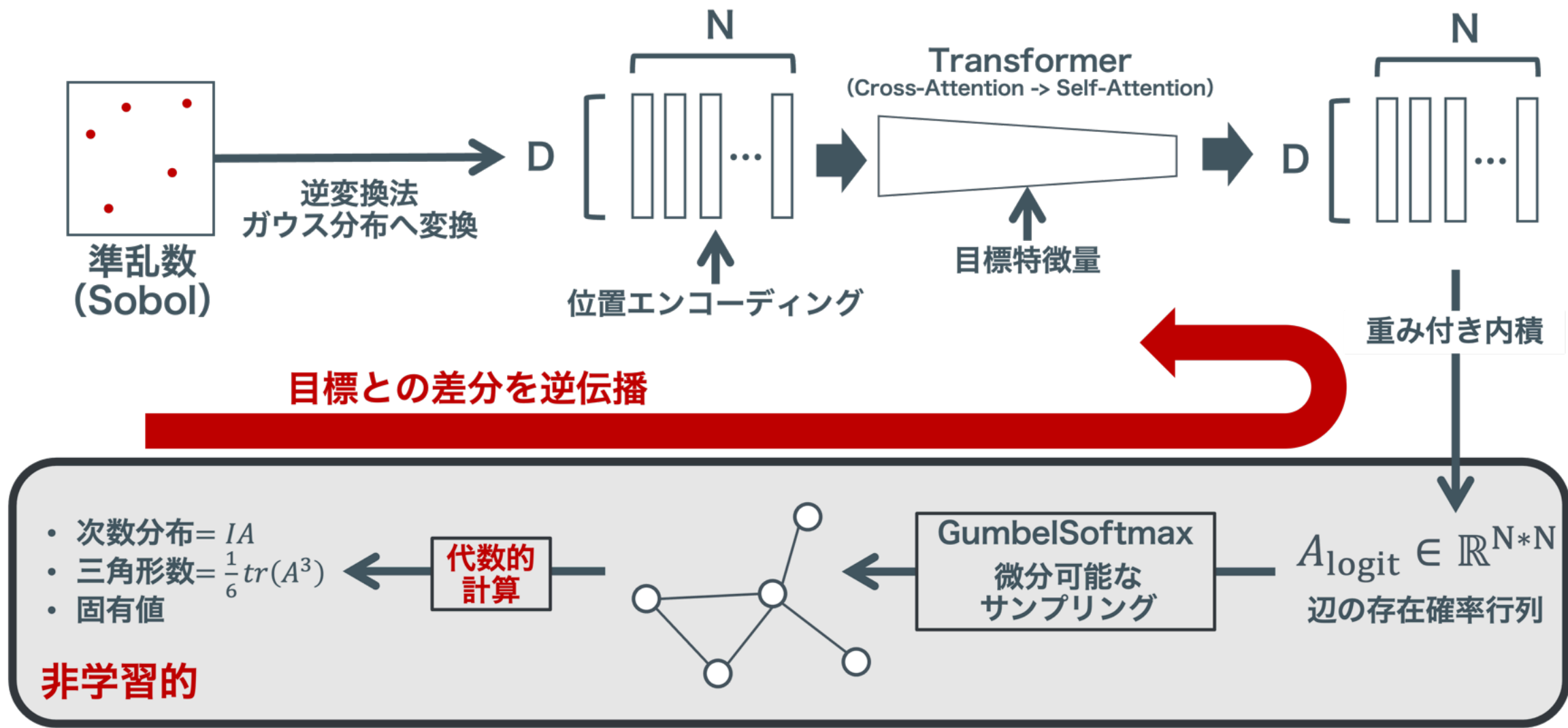
目標APL=4の生成グラフをWLハッシュで数えたユニーク構造数の分布で比較する。



WKDEは目標付近への集中度が高い一方、ユニークな構造数は325→176に減少する。

DATALESS / TRAINING-FREE

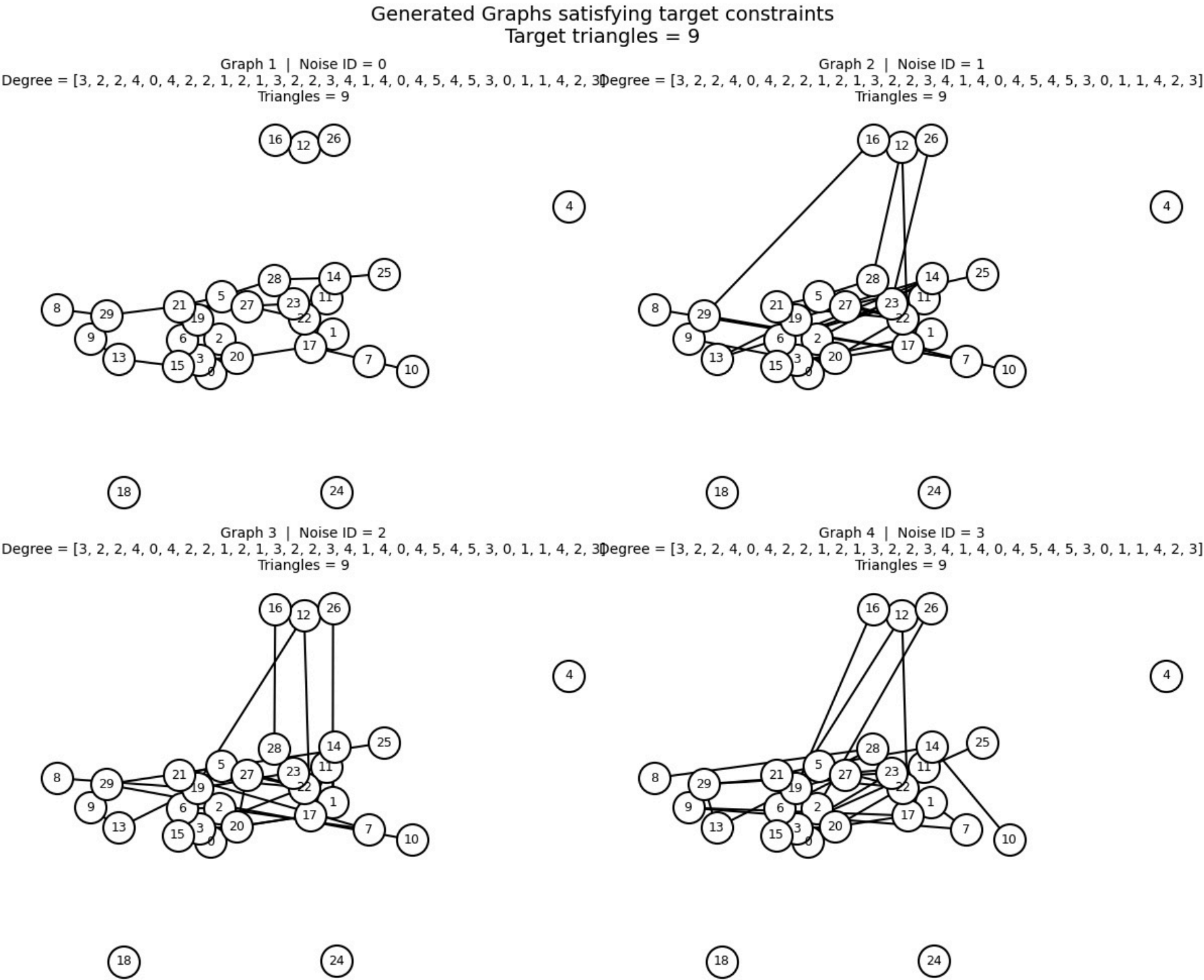
データレス・非学習グラフ生成



GENERATION RESULT

三角形数9を目標にした30ノードグラフの生成例

同じ次数列のままノイズを変えた4通りの生成結果すべてが目標の三角形数に一致する。

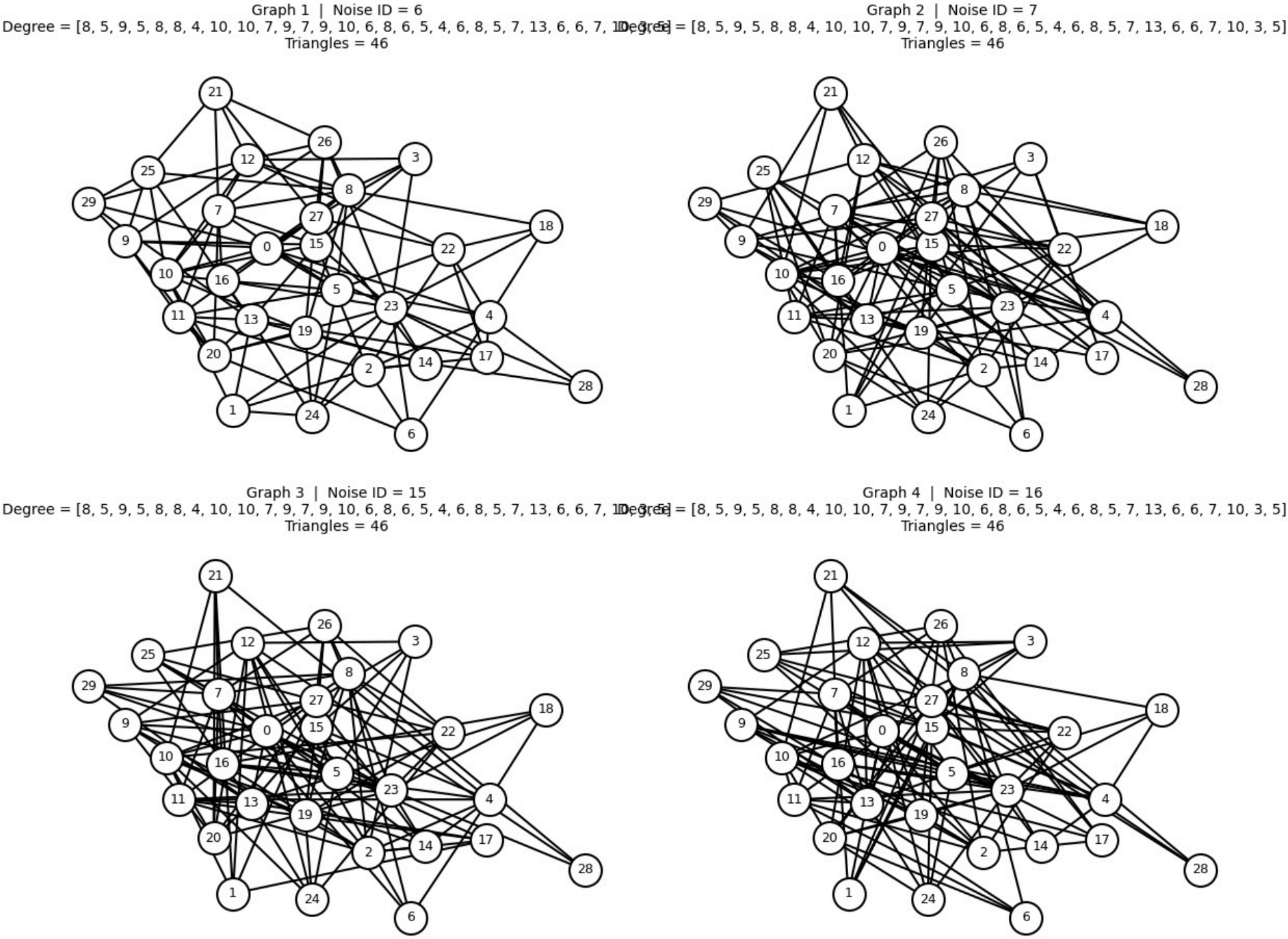


GENERATION RESULT

三角形数46を目標にした30ノードグラフの生成例

目標値を大きく引き上げても、**同じ枠組みでより密なグラフ**を正確に生成できる。

Generated Graphs satisfying target constraints
Target triangles = 46



SIMPLICITY

モデルのシンプルさの追求と展望

ランダム写像によるモデル

Transformerは表現力が過剰かもしれない。ランダム写像による代替を検討する

代数的制約の拡充

扱える制約の種類と組合せを増やす

ルールベースの事後制約

生成後のグラフにルールを適用し、制約を満たすように補正する

網羅性の検証

制約を満たすグラフ集合をどの程度網羅できるかを検証する

SUMMARY

ネットワークにおけるまとめ — 再現、制御、保証

再現 実データの統計的性質を再現
入手困難な通信トラフィック・グラフを補う

制御 DiffuPac・GraphTuneで目標特性へ誘導
出力のばらつきと、精度・多様性の両立が課題

保証 非学習的で目標条件の充足へ
30ノードで次数列・三角形数が目標に一致する例を確認

次の課題 事後制約の導入・代数的制約の拡充
ランダム写像による簡素化・網羅性の検証

ご清聴ありがとうございました

Questions & Discussion