修士論文 要旨

二次マルコフ過程を考慮した自己回帰型テンポラルグラフ生成モデルと送金データへの応用

数理情報学専攻 48216223　長澤 達也
指導教員　久野 遼平 講師

1　はじめに

送金データとは口座やユーザー間の送金履歴を記録したデータのことである．これらはネットワークとして捉えることができ，経済における資金の流れをリアルタイムで追うことができる貴重なデータである．国内の都市銀行から提供された銀行送金データに対する実証分析 [3] では，送金ネットワークがスケール性やスモールワールド性といった複雑なネットワークの特性を持つことや，経済環境の変化を色濃く反映したデータであること，着金元の情報が送金先に相関するという二次マルコフ過程の性質を持つことなどが明らかになった．

そこで本修士論文では送金データを対象にした動的・有向・中規模ネットワークのテンポラルグラフ生成を試みる．グラフ生成モデルとは訓練対象となるネットワークの集合を対象に，訓練対象と同じような統計的性質を持つネットワークを生成するアプローチのことである．深層学習を用いたグラフ生成は既に分子グラフなどにおいて既に一定の成果を挙げているが，有向ネットワークを対象にしたものや時間変化を伴うネットワークに対してはまだ研究が浅い．

本修士論文では実証分析の結果を踏襲し [3]，送金ネットワークにおけるテンポラルグラフ生成問題を設定し，モデルを提案する．提案モデルは自己回帰型グラフ生成モデルをベースに時間拡張を行ったうえで，実証分析の結果得られた一月前の着金元が当月の送金先と相関するという性質を加味したものである．

2　問題設定

時刻 t における有向グラフ $G^t = (V^t, E^t)$ をノード集合 $V^t = \{v_1, \ldots, v_{N^t}\}$ およびエッジ集合 $E^t = \{e_{uv} = (u, v) | u, v \in V^t\}$ によって定義する．各時刻における訓練グラフの集合列:

$$\mathcal{D}_{\text{train}} = (\mathcal{G}^1, \ldots, \mathcal{G}^T | \mathcal{G}^t = \{G_i^t\}) \tag{1}$$

が与えられたもとで，$p(G^t)$ を予測し，G^{T+1} を生成することが本モデルの目標である．さらに，本モデルは時系列データを対象としているからその時刻までのグラフを条件として用いることができ，$p(G^t)$ は $p(G^t|G^{<t})$ と表すことができる．

3　提案モデル

本モデルは静的自己回帰型グラフ生成モデルである GraphGen[1] をベースに時間拡張を行ったうえで，対象を無向グラフから有向グラフに拡張したものである．モデルの全体の流れを図 1 に示す．

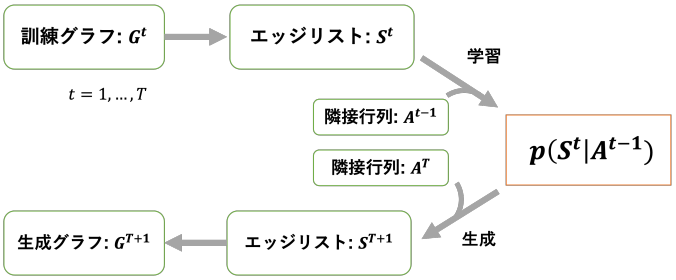

図 1. 提案モデルの全体図

まず，グラフに対し登場順によってノードの順番を定め，これを用いて辞書順で並び替えたエッジリスト S^t に変換する．ここでエッジリスト $S^t = (s_1, \ldots, s_m)$ の構成要素 s_i は

$$\begin{aligned} s_i &= (t_u, t_v, L_u, L_e, L_v) \\ t_u, t_v &: \text{ノード } u, v \text{ のノード番号} \\ L_u, L_v &: \text{ノード } u, v \text{ のラベル情報} \\ L_e &: \text{エッジ } e_{uv} \text{のラベル情報} \end{aligned} \tag{2}$$

と 5 つの成分からなるタプルとして定義する．

二次マルコフ過程の考慮として，このエッジリストに対して 1 時点前の隣接行列 A^{t-1} を条件として用いる．すなわち，本モデルの条件付き生成モデルは

$$p(G^t|G^{<t}) = p(S^t|A^{t-1}) \tag{3}$$

と表せる．LSTM を用いてエッジを順に生成するので，

$$p(S^t|A^{t-1}) = \prod_{i=1}^{m} p(s_i|s_{<i}, A^{t-1}) \tag{4}$$

$$= \prod_{i=1}^{m} p((t_u, t_v, L_u, L_e, L_v)|s_{<i}, A^{t-1}) \tag{5}$$

となる．隣接行列 A^{t-1} の u に対応する列 A_u^{t-1} は u の 1 時点前の着金元情報を表しており，これを送金先の予

測 (t_v, L_v, L_e) に条件として用いる．さらに本モデルではエッジの始点ノードを表す t_u，L_u も送金先の条件に与える．すなわち，

$$p(S^t|A^{t-1}) = \prod_{i=1}^{m} p((t_u, L_u)|s_{<i}) p((t_v, L_v, L_e)|s_{<i}, (t_u, L_u), A_u^{t-1}) \tag{6}$$

と表せる．(t_u, L_u) と (t_v, L_v, L_e) の組はそれぞれ独立性を仮定し，それぞれの条件を LSTM の隠れ状態 h_i^1, h_i^2 を用いて表すと

$$p(S^t|A^{t-1}) = \prod_{i=1}^{m} p(t_u|h_i^1) \cdot p(L_u|h_i^1) \cdot p(t_v|h_i^2) \cdot p(L_v|h_i^2) \cdot p(L_e|h_i^2) \tag{7}$$

である．h_i^1, h_i^2 は埋め込み関数 f_{emb} と LSTM の状態遷移関数 f_h を用いて

$$h_i^1 = f_h\left(h_{i-1}^1, f_{\mathrm{emb}}(s_{i-1})\right) \tag{8}$$

$$h_i^2 = f_h\left(h_{i-1}^2, f_{\mathrm{emb}}(s_{i-1}, (t_u, L_u), A_u^{t-1})\right) \tag{9}$$

で更新される．また，各成分 $c \in \{t_u, t_v, L_u, L_e, L_v\}$ は対応する h_i と出力関数 f_c を用いて

$$c \sim_M \theta_c = f_c(h_i) \tag{10}$$

によって得る．θ_c は多項分布のパラメータであり，$\sim_M$ は多項分布によるサンプリングを表す．

以上のエッジ s_i を生成する一連の過程を図 2 に示す．

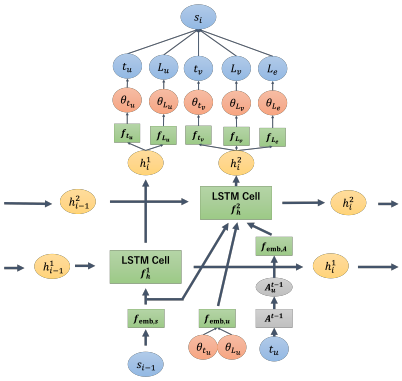

図 2. 提案モデルのネットワーク図

4　主結果

提案モデル，GraphGen および古典的ネットワークモデルについて，銀行送金データに対してグラフ生成した結果の一部を表 1 に示す．

いずれのデータにおいてもブロック構造を用いた古典的ネットワークモデルである DCSBM が優れた値を示しており，深層グラフ生成モデルである提案手法および GraphGen は送金ネットワークのグラフ構造を学習できていないと言える．提案手法と GraphGen の間では優位な差は見られない．

古典的ネットワークモデルではラベル推定を行わないため，ラベルの類似性は提案手法と GraphGen のみ評価を行う．ノードラベルでは GraphGen の方が高精度を示した一方で，エッジラベルでは提案手法の方が優れた結果を出している．エッジラベルの予測に対し，提案モデルではそれまでのエッジ情報に加え，始点ノードの情報 (t_u, L_u) および A_u^t を与えていることが精度の向上に繋がった可能性がある．

表 1. 生成グラフに対するグラフ統計量の指標．各指標の値は分布に対する MMD[2]．数値が小さいほどテストグラフと生成グラフが類似していることを表す．

モデル	総次数	入次数	出次数	クラスタ係数	NSPDK	ノードラベル	エッジラベル
提案モデル	0.75	1.12	1.02	1.6961	0.1637	0.2247	**0.0011**
GraphGen	1.00	1.22	1.24	1.8195	0.1396	**0.1305**	0.0018
Erdös	0.33	0.64	0.67	1.5935	**0.1021**	-	-
BA	0.90	1.10	1.12	1.7376	0.2755	-	-
SBM	0.20	0.35	0.35	0.7696	0.1035	-	-
DCSBM	**0.15**	**0.29**	**0.26**	**0.7509**	0.1033	-	-

5　おわりに

本修士論文では送金データを対象とした動的・有向・中規模ネットワークのテンポラルグラフ生成モデルの構築を試みた．結果としては提案モデルを目覚ましく改善する所まではもっていくことができなかった．しかし，送金ネットワークに対するテンポラルグラフ生成問題の有効性を示すことができた．また同時に送金ネットワークに対してテンポラルグラフ生成問題を設定する際に生じる諸問題を整理することができた．

参考文献

[1] Nikhil Goyal, Harsh Vardhan Jain, and Sayan Ranu. Graphgen: A scalable approach to domain-agnostic labeled graph generation. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pp. 1253–1263. ACM / IW3C2, 2020.

[2] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, Vol. 13, No. 1, pp. 723–773, 2012.

[3] 久野遼平, 長澤達也, 高橋秀, 近藤亮磨, 大西立顕. 銀行送金ネットワークの内在的構造と時間変化. 人工知能, Vol. 38, No. 2, 2023.