

厳密な期待値計算に基づく Uncertain グラフにおけるロールクラスタリング

内藤 綜志† 伏見 卓恭†

†東京工科大学・コンピュータサイエンス学部

1 はじめに

ネットワークサイエンスの分野において、有向3ノード Motif における構造同値性に基づき各ノードの役割を定義した Role が提案され、それをを用いた分析結果が報告されている [1, 2]. 図 1 は、グラフ同型に基づき有向3ノードサブグラフを 13 パターンに分類したものであり、Motif 1~Motif 13 として定義する. 同様に、構造同値性に基づきサブグラフ内のノードを 30 パターンに分類したものであり、Role 1~Role 30 として定義する. この Role を利用することで、例えば、Role 13 として出現したノードは情報を発信することに特化した役割を担っており、また一方で Role 24 として出現したノードは受け取った情報を他ノードに発信するといった役割を担っていると特徴づけることができ、バイラルマーケティングにおいて重要なインフルエンサーなどの特定が可能になると期待できる.

また、ここ数年のネットワークサイエンスにおけるトレンドとして Uncertain グラフの研究がある [3]. Uncertain グラフは、エッジに対して出現確率が付与されたもので、通常のグラフと比較してより現実世界をより精緻に表現することが可能である. しかし、エッジの存在が不確実なままでは、上述した Role の出現回数を確定的にカウントできないので、不確実エッジ数を L としたとき、 2^L 個の起こりうるグラフを全て列挙し、それらに対して Role 数をカウントして、各グラフの

生起確率による重み付きの平均値を計算する必要がある. 通常は、一定数のグラフサンプリングを行い、各サンプルグラフから Motif または Role をカウントし、それらをサンプル数で割ることで期待値の近似値を求める [4, 5]. サンプル数を増やすことで、期待値を精度よく近似できるが、サンプル数に比例して計算量が増加してしまうという問題点がある. そこで本研究では、サンプリングすることなく厳密な期待値を計算する手法を提案し、Uncertain グラフにおける Role カウントを行う. 実験では、精度と実行時間の観点から、既存サンプリング手法と比較し提案法の有用性を示す.

2 提案手法

ノード集合 V とエッジ集合 E からなるグラフ構造 $G = (V, E)$ において、エッジの存在が確率的に決定するグラフを Uncertain グラフ $\mathcal{G} = (G, p)$ と呼ぶ. ここで、 G はバックボーングラフ、 $p: E \rightarrow (0, 1]$ はエッジ存在確率である. \mathcal{G} の各ノードに対して、図 1 に示す 30 種類の Role として出現する頻度の期待値を求める問題を扱う. バックボーングラフ G におけるノード $v \in V$ の Role i としての出現頻度を $R_{v,i}$ と表記する. Role i を含む Motif のエッジ数を $l(i)$ と表記すると、全てのエッジの存在確率が同じ値の場合、すなわち、 $p(e) = p$ の場合、Role i の期待値 $R_{v,i}(p)$ は、 $p^{l(i)} \cdot R_{v,i}$ である. 例えば、 G においてノード v が Role 30 として $R_{v,30}$ 回出現したとする. Role 30 を含む Motif 13 のエッジ数は $l(30) = 6$ であるため、6 本のエッジ全てが出現する確率は p^6 であり、ノード v が Role 30 として出現する頻度の期待値は $R_{v,30}(p) = p^6 R_{v,30}$ となる. 一方、 $1 - p^6$ の確率で Motif 13 は崩壊し他の Motif となるため、ノード v は他の Role として出現することになる. 具体的には、6 本中 1 本のエッジが出現しないことを表す $p^5(1-p)$ の確率で Motif 12 における Role 29 へと変化したが、ノード v が Role 30 から 29 へと変化するために必要な非出現エッジは 2 パターンあるため、確率は $2p^5(1-p)$ となる. また、バックボーングラフにおいて Role 29 として出現していた頻度も考慮すると、ノード v が Role 29 として出現する頻度の期待値は $R_{v,29}(p) = 2p^5(1-p)R_{v,30} + p^5R_{v,29}$ となる. 同様に、他の Role に関してもエッジ存在確率に従って別の Role へと変化する. これらを一般化する

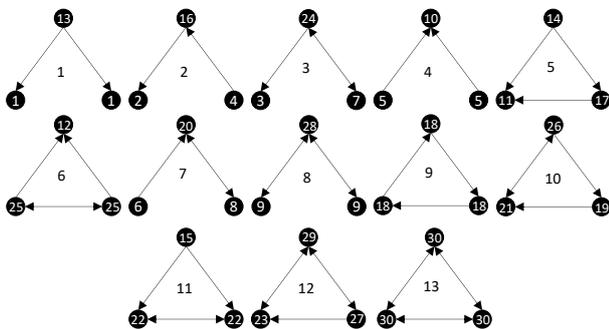


図 1: 有向3ノードからなる 13 Motif における 30 Role

Role Clustering in Uncertain Graphs Based on Exact Expected Value Calculations
 †Soshi Naito †Takayasu Fushimi
 †School of Computer Science, Tokyo University of Technology

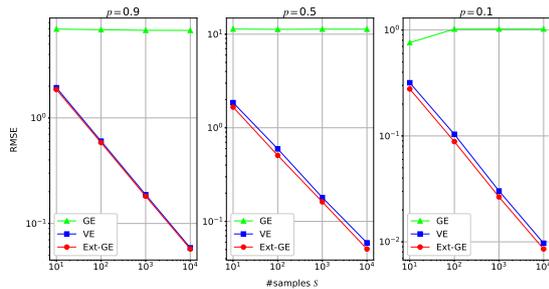


図 2: 既存手法の RMSE

と、ノード v が Role i として出現する頻度の期待値は

$$R_{v,i}(p) = \sum_{j=1}^{30} T(j \rightarrow i) p^{l(i)} (1-p)^{(j)-l(i)} R_{v,j}$$

と書ける。したがって、バックボーングラフ G における全 N ノードの Role 出現頻度ベクトルを並べた $(N \times 30)$ の行列 $\mathbf{R} = [\mathbf{r}_1, \dots, \mathbf{r}_N]^T$ に対して、Role 間の推移確率行列 \mathbf{T} とエッジ出現確率行列 \mathbf{P} 、非出現確率行列 \mathbf{Q} のアダマール積により得られる行列を掛けることで、Role 数の出現頻度に関する厳密な期待値をサンプリングなしに高速に計算することができる：

$$\mathbf{R}(p) = \mathbf{R}(\mathbf{T} \otimes \mathbf{P} \otimes \mathbf{Q}).$$

紙面の都合上、推移確率行列の具体的な値は割愛する。

3 評価実験

評価実験では、複数の実有向グラフに対して、一様なエッジ出現確率 $p(e) = p \in \{0.1, 0.5, 0.9\}$ を付与したものをを用いる。それらに対して、提案手法およびサンプリングに基づく既存手法 [5] により有向 3 ノードモチーフをもとにした Motif-Role を抽出し、誤差と実行時間の観点から評価する。サンプル数は $S \in \{10^1, 10^2, 10^3, 10^4\}$ とした。紙面の都合上、Blog データの結果のみを示す。

図 2 は、サンプリングに基づく既存手法による近似値 $\hat{\mathbf{R}}(p)$ と提案手法による厳密な期待値 $\mathbf{R}(p)$ との Root Mean Squared Error (RMSE) である。横軸はサンプル数、縦軸は対数スケールで表した RMSE である。この結果から、サンプリングに基づく手法は、いずれもサンプル数が増えるほど RMSE が小さくなるものの、一定の誤差が生じてしまっていることがわかる。

図 3 は、各種法がクラスタリング結果を出力するまでの実行時間を示す。横軸はサンプル数、縦軸は対数スケールで表した実行時間である。提案手法を Prop と表記した。この結果から、サンプル数に依存しない提案手法は、 $S = 10^2$ 個のサンプルを用いる既存手法と同程度の速度であることが分かる。これらの結果から、

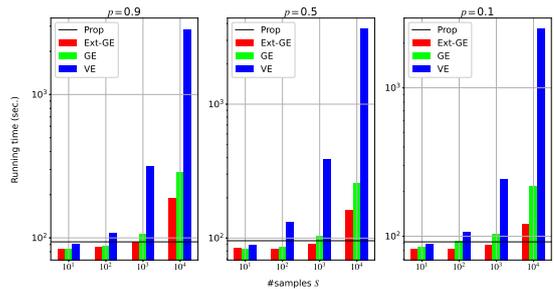


図 3: 提案手法と既存手法の実行時間

以下のことが確認できる。1) 既存手法はサンプル数が増えるごとに誤差は小さくなり、精度が高くなる。2) 既存手法はサンプル数が増えるごとに実行時間が長くなる。3) 提案法は厳密な期待値を計算しているため誤差はなく、実行時間は一定で、非常に高速である。

4 おわりに

本研究では、Uncertain グラフから Role カウントを行うタスクにおいて、既存の手法における問題点を指摘し、それらを解消するサンプリングを行わない厳密な期待値計算法を提案した。実験では、実ネットワークに一樣なエッジ確率を付与したものをを用い、既存の手法と比較して、精度と実行時間の面で評価した。結果として、提案法が最先端のアルゴリズムを取り入れた手法や著者らが以前提案した手法と比較して圧倒的な速度で厳密な期待値を出力できることを確認した。

今後の課題としては、エッジの出現確率を非一様に付与したネットワークでの実験が挙げられる。

謝辞 本研究は、JSPS 科研費 (No.20K11940) (No.19K20417) の助成を受けたものである。

参考文献

- [1] T. Ohnishi, H. Takayasu, and M. Takayasu. Network motifs in an inter-firm network. *Journal of Economic Interaction and Coordination*, Vol. 5, No. 2, pp. 171–180, 2010.
- [2] M. D. McDonnell, O. N. Yaveroglu, B. A. Schmerl, N. Iannella, and L. M. Ward. Motif-role-fingerprints: The building-blocks of motifs, clustering-coefficients and transitivities in directed networks. *PLOS ONE*, Vol. 9, No. 12, pp. 1–25, 12 2014.
- [3] A. Khan, Y. Ye, L. Chen, and H. V. Jagadish. *On Uncertain Graphs*. Morgan & Claypool, 2018.
- [4] C. Ma, R. Cheng, L. V. S. Lakshmanan, T. Grubenmann, Y. Fang, and X. Li. Linc: A motif counting algorithm for uncertain graphs. *Proc. VLDB Endow.*, Vol. 13, No. 2, p. 155–168, October 2019.
- [5] S. Naito and T. Fushimi. Motif-role extraction in uncertain graph based on efficient ensembles. In *Proceedings of the 10th International Conference on Complex Networks and Their Applications*, 2021.