DEIM Forum 2022 G34-4

拡張グラフアンサンブル法に基づくUncertainグラフにおけるMotif-Role抽出の精度向上

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　〒192–0982 東京都八王子市片倉町１４０４－１
E-mail: †{c011833449,fushimity}@edu.teu.ac.jp

あらまし　本稿では，エッジの存在が確率的に決定されるUncertainグラフから，Motifを拡張した概念であるMotif-Roleを抽出する問題に取り組む．Uncertainグラフから Motif-Roleの抽出を高精度で行うには，多数回のグラフサンプリングをし，各サンプルグラフに対して，Role数カウント，類似度計算，クラスタリングが必要になる．この問題に対して，著者らはグラフアンサンブル法を提案し，Uncertainグラフの Motifカウントに関する最先端技術を用いたベクトルアンサンブル法と類似度アンサンブル法と比較して，最も高速に類似した結果を出力することを確認した．しかし，真の値との誤差を比較した時，グラフアンサンブル法はベクトルアンサンブル法より誤差が大きいことがわかった．そこで，本研究ではグラフアンサンブル法を拡張し，効率性を維持したまま，誤差を小さくする新たな手法を提案する．比較実験では，提案手法を真の値との誤差，効率性の観点から評価する．実験データには各エッジに一様な出現確率を付与した3つの実ネットワークを用い，提案法が他の手法と比較して，真の値との誤差が少ないまま同程度の速度で出力できることを示す.

キーワード　Uncertainグラフ，　Motif-Role

1　はじめに

ネットワークサイエンスの分野において，小規模なサブグラフである Motifの数をカウントすることは，そのグラフの特徴を理解する上で重要なタスクであり，Miloらの研究 [1] に端を発し長年に渡って研究されてきた [2], [3], [4], [5], [6]．Motifの概念を拡張する研究は盛んに行われており，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 Motif-Roleが提案され，それを用いた分析結果が報告されている [7], [8]．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，Motif1〜Motif13と定義する．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を Role1〜Role30と定義する．この Motifに基づく Roleを利用することで，例えば，Role13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，Role24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る．このように，ノードごとの Motifにおける役割を抽出することで，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（不確実グラフ）の研究がある．Uncertainグラフは，各エッジに出現確率が定義されたもので，道路ネットワークにおける通行止めや SNSにおけるユーザーのフォロー/フォロワー関係など，現実世界におけるエッジの存在の不確実性を表現することが可能である．Uncertainグラフでの Motifや Roleのカウントは，現実に即したグラフの詳細な分析を行うことができるため，マーケティングをはじめとし，都市計画，タンパク質の分析といった幅広い分野での活用が期待される．しかし，エッジの存在が不確実なままでは，上述した Motifや Roleの出現回数を確定的にカウントすることができないため，このような確率グラフに対しては，Motifや Roleの出現回数の期待値を考えることになる．L 本の不確実エッジを持つ Uncertainグラフに対して正確な期待値を求めるには，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 Motifまたは Roleの数をカウントして，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．しかし，小規模なグラフであっても組み合わせ数 2^L が非常に大きくなるため，一般に計算は困難である．そのため，一定数のサンプリングによる近似が一般的である．

Uncertainグラフにおける Motifのカウントに関する既存研究として，Maらによる LINCアルゴリズム [9] がある．このアルゴリズムは，全サンプルグラフに対して1から Motifをカウントするのではなく，サンプルグラフ間の構造類似性に着目し，2サンプルグラフ間の差分エッジおよびその周辺のみに焦点を当てることで，効率的に Motif数を更新する．LINCアルゴリズムは Uncertainグラフ上での Motifカウントアルゴリズムとして最先端のものであり，エッジの平均出現確率が高いグラフにおいてナイーブなサンプリングリング手法よりも高速に結果を出力できるが，不確実度が高い Uncertainグラフ，すなわち，エッジの平均出現確率が 0.5に近いグラフや，Uncertainエッジの数が多いグラフでは有効ではない．

本稿では，Motifに基づく Roleの抽出問題を定式化し，Uncertainグラフのサンプルに対する効率的なアンサンブルにより Roleを抽出することを試みる．既存手法として，Uncertain

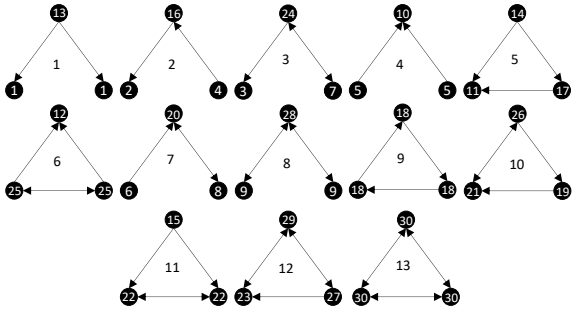


図 1　有向 3 ノード Motif における 30 Role

グラフにおける Motif カウントの最先端技術である LINC アルゴリズムを利用し，多数のサンプルグラフから Role 数を効率的にカウントするベクトルアンサンブル法とそれを応用した類似度アンサンブル法が存在する．著者らは，サンプリングしたグラフをアンサンブルし，各エッジに全サンプル中での出現回数を重みとした重み付きグラフに対して Role を抽出するグラフアンサンブル法を提案した [10]．そして，グラフアンサンブル法とベクトルアンサンブル法，類似度アンサンブル法の出力結果が互いに比較的類似すること，グラフアンサンブル法が他の 2 手法と比較して圧倒的に高速に結果を出力することを確認した．しかし，各ノードにおける Role 数の期待出現頻度の厳密値（真の値）に対する誤差の観点で，グラフアンサンブル法はベクトルアンサンブル法に劣っていることがわかった．これは，重み付きグラフであるアンサンブルグラフにおいて Role 数をカウントするだけでは，確率的に生じるエッジの不在による Motif の崩壊を考慮できていないからである．たとえば，Motif 13 を構成する 6 本のエッジの存在確率が p であったとすると，Uncertain グラフにおいても Motif 13 および Role 30 である確率は p^6 である．すなわち，$1-p^6$ の確率で他の Motif および Role に変化する．

本研究では，このことを考慮した新たなグラフアンサンブル法を提案し，グラフアンサンブル法の長所である効率性を維持したまま，最先端アルゴリズムである LINC を用いたベクトルアンサンブル法に匹敵する精度を実現することを目指す．

2　関 連 研 究

2.1　Motif カウント，Role 抽出

Milo らによる先駆的な研究 [1] をきっかけに，モチーフのカウントの技術は何年にもわたって研究され，多くのアルゴリズムが多用な目的のために開発されてきた [3]．Itzhack らは，ターゲットノードをルートとする幅優先探索木をトラバースする効率的なアルゴリズムを提案した．これは，サブグラフ内のリンクの存在をビット文字列として表し，各サブグラフの同型性を調べることなくモチーフパターンを効率的に識別できる [3]．本研究においても，サンプルグラフからのモチーフカウントに，Itzhack らのアルゴリズムを採用する．また，McDonnell らは有向 3 ノードの構造同値性に着目することでモチーフロールを定義し，モチーフの出現回数を表す行列をモチーフロールの出現回数を表す行列へと変換する変換行列を提案した [8]．本論文で対象とするモチーフロールは McDonnell らにより定義されたモチーフロールと同じものであるが，カウンティングに使用するアルゴリズムは McDonnell らのものではなく，前述した Itzhack らによるものを用いる．また，モチーフロールの集計結果を応用した研究として，Sarajlić らの研究がある．この研究では独自に定義したモチーフロールを用い，ノードが国，エッジが取引を表す世界の貿易情報が記されたネットワークの分析手法を提案し．ここから，ネットワーク中での各国の特徴分析，各国の経済状況の予測，国同士の関係性の分析，各役割の特徴分析などを実験，評価している [11]．

2.2　Uncertain グラフ

Uncertain グラフに関する研究は信頼性・クエリ・マイニングなど広いコンテキストで研究されてきており，モチーフカウントを含む既存のグラフ分析手法を Uncertain グラフに拡張することは重要なタスクとなっている．しかし，Uncertain グラフに対するモチーフカウントの研究はまだ十分にされておらず，より多くの研究が行われることが期待されている．今回は既存研究の中でも，主要なものを以下にあげていく．Hu らは，単純なエンベディングを用いた URGE という方法を提案した．これはグラフ中の各ノードの近接情報を持つ低次元ベクトルの集合を生成することでノードの近接情報を保持したままグラフの次元を小さくすることができる．これにより，Uncertain グラフの複雑性が高い，既存のグラフマイニングのアルゴリズムが適用できないといった Uncertain グラフを扱う上で直面する問題に対応することが可能になる [12]．Ma らは，サンプリングに基づいて，モチーフ数の平均，分散，確率分布などの基本的な統計を求める 2 つのアルゴリズムを提案した [9]．1 つ目は，PGS と呼ばれる単純なサンプリング方法で，Uncertain グラフから多数の実現グラフをサンプリングし，各サンプルグラフから単一のモチーフのインスタンスをカウントする．さらに，Hoeffding の不等式に基づいてモチーフ数の平均を正確に推定するのに十分なサンプル数について議論している．2 つ目は，LINC と呼ばれるより効率的な方法で，サンプルグラフ間の構造的類似性を利用し，連続するサンプル間のエッジの違いのみを調べ，モチーフの頻度を更新する．LINC は，全く同じサンプルグラフを使用した場合，PGS と同じ結果を出力するが，PGS よりもはるかに高速に動作するアルゴリズムである．しかし，エッジの出現確率が低いものを多く含むグラフや，エッジの数が多いグラフになると実行時間が PGS より遅くなってしまうという課題も存在する．本研究では，LINC アルゴリズムを最先端の技術と見なす．

3　問 題 設 定

本研究では，Uncertain グラフに対して，Motif-Role の観点から各ノードが果たす役割を抽出する問題を扱う．説明の便宜上，有向 3 ノードの Motif に基づく Role 抽出の場合について説明するが，有向，無向に限らず小規模な k ノード Motif に基

づく Role 抽出に適用できる．

3.1 Motif-Role 抽出

はじめに，バックボーングラフ $G=(V,E)$ から Motif-Role を抽出する問題を定式化する．ここで，V はノード集合，E はエッジ集合，$N=|V|$ はノード数，$L=|E|$ はエッジ数である．McDonnell らの研究 [8] に従い，Motif における構造的同値性に基づき Role を定義する．有向 3 ノード Motif においては，図 1 に示す 30 種類の Role が存在する．本研究における

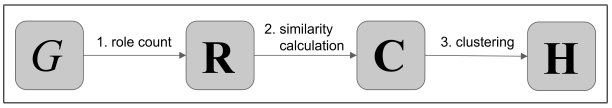


図 2　Motif-Role の抽出手順

Motif-Role 抽出は，1) ノードごとの Role ベクトルの構築，2) 全ノードペアに対する Role ベクトル間の類似度計算，3) 類似度に基づくクラスタリングの 3 ステップにより類似 Role を有するノードグループを抽出する（図．2）．手順 1 のノードごとの Role ベクトルの構築では，各ノード v に対して R 個の Role の数をカウントし，その出現頻度を並べた R 次元ベクトル $\mathbf{r}_v$ を構築する．$\mathbf{r}_v$ における i 番目の要素は，ノード v が Role i のとして出現した回数を表している．有向 3 ノード Motif の場合，図 1 に示さように $R=30$ である．全 N ノードの Role ベクトルを並べた行列を $\mathbf{R}=[\mathbf{r}_1,\ldots,\mathbf{r}_N]^T$ と表記する．ここで $\mathbf{r}^T$ は $\mathbf{r}$ の転置を表す．手順 2 では，Role ベクトル間のコサイン類似度 $c_{u,v}=\frac{\mathbf{r}_u^T\mathbf{r}_v}{||\mathbf{r}_u||||\mathbf{r}_v||}$ を用い，全てのノードペアの類似度を計算する．全 $N\times N$ のノードペアの類似度を類似度行列 $\mathbf{C}=[c_{u,v}]_{u\in V,v\in V}$ とする．手順 3 では，全てのノードを k-medoids クラスタリングの貪欲法 [13] により K 個のクラスタに分類する．すなわち，もしノード u がクラスタ k に所属している場合，$h_{u,k}=1$ そうでない場合は 0 とする．$h_{u,k}\in\{0,1\}$ はノード u がクラスタ k に所属している情報，$\mathbf{h}_u=[h_{u,k}]_{k=1}^K$ は所属ベクトルとなり，$\mathbf{H}=[\mathbf{h}_1,\ldots,\mathbf{h}_N]^T$ が出力となる．結果としては，類似の Role ベクトルを有するノードクラスタが得られる．この一連のプロセスを Role 抽出と定義する．

3.2 Uncertain グラフ

本研究では，ノード間のエッジが存在するか否かが確率的に決定する Uncertain グラフを対象とする．Uncertain グラフ $\mathcal{G}=(G,p)$ は，ノード集合 V とエッジ集合 E からなるバックボーングラフ $G=(V,E)$ と各エッジの存在確率 $p:E\to(0,1]$ により定義される．Uncertain グラフは，その実現可能グラフの集合としても表現できるため，$\mathcal{G}=\{G_i=(V,E_i);E_i\subseteq E\}$ と表す．Uncertain エッジの数を L とすると，Uncertain グラフの実現可能グラフの数は $2^L=|\mathcal{G}|$ である．関連研究に倣い，各実現可能グラフ G_i の生起確率 $\Pr[G_i]$ をすべてのエッジに対する独立ベルヌーイ試行に基づき計算する：

$$\Pr[G_i]=\prod_{e\in E_i}p(e)\prod_{e\in E\setminus E_i}(1-p(e)).$$

3.3 Uncertain グラフでの Motif-Role 抽出

次に，Uncertain グラフ $\mathcal{G}$ から Motif-Role を抽出する問題を定式化する．Uncertain グラフに対して上述の Role 抽出問題を厳密に解くには，全ての実現可能グラフ $\mathcal{G}=\{G_i=(V,E_i);E_i\subseteq E\}$ に対して上述の 3 つの手順を行い，それらの結果を各実現可能グラフ G_i の生起確率 $\Pr[G_i]$ を考慮してアンサンブルする必要がある．そして，以下のように各実現可能グラフに対する Role 抽出結果をアンサンブルする：

$$\mathcal{H}=\bigoplus_{G\in\mathcal{G}}(\mathbf{H}_G;\ \Pr[G]).$$

ここで，Φ はアンサンブルの演算子であり，各実現可能グラフ G のクラスタリング結果 $\mathbf{H}_G$ が生起確率 $\Pr[G]$ の重みを考慮してアンサンブルされていることを表す．L 本の Uncertain エッジを有する Uncertain グラフに対して厳密なアンサンブル結果を得るには，実現グラフの数 $|\mathcal{G}|=2^L$ だけサンプリングとアンサンブルが必要になり，小規模なグラフに対しても計算は困難であるため，サンプリングによる近似が一般的に採用される．

4 既存手法

この節では，Uncertain グラフから実現可能グラフをサンプリングし，クラスタリング結果を出力する 4 つのアンサンブル手法について説明する（図 3 参照）．

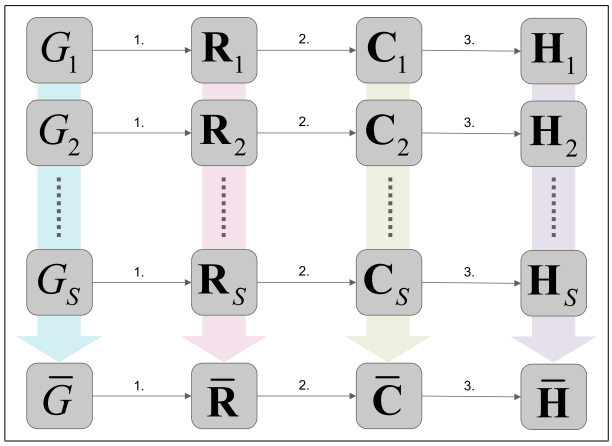


図 3　4 つのアンサンブル法

4.1 グラフアンサンブル法

まず，著者らによって提案されたグラフアンサンブル法について説明する [10]．グラフアンサンブル法（以下，GE 法）において類似度行列 $\bar{C}$ を出力するまでの手順を Algorithm 1 に示す．GE 法は，サンプルグラフ群 $\{G_1,\ldots,G_S\}$, $G_s=(V,E_s)$, $E_s\subseteq E$ に対しアンサンブルを行い，アンサンブルグラフ $\bar{G}$ を生成する（Algorithm 2）：

$$\mathcal{G}=\bigoplus_{G\in\mathcal{G}}(G;\ \Pr[G])\simeq\bigoplus_{s=1}^{S}(G_s;\ 1/S)=\bar{G}.$$

$\bar{G}=(V,\bar{E},\bar{p})$ は S 回のサンプルの中でエッジ e が出現する確率 $\bar{p}(e)=\sum_{s=1}^{S}\delta(e\in E_s)/S$ を重みとした重み付きグラフで

Algorithm 1 グラフアンサンブル法：GE($\mathcal{G}, S$)

Input: $\mathcal{G} = (G, p),\ G = (V, E),\ \ S$
Output: $\bar{\mathbf{C}}$
Initialize: $\bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leqq i \leqq 30} \leftarrow \mathbf{0}$
$\bar{G} \leftarrow$ emsemble_graphs($\mathcal{G}, S$)
$\Psi \leftarrow$ search_connected_triples($\bar{G}$)
for $G^{(m)} \in \Psi$ **do**
    $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)})$
    $\bar{\mathbf{R}} \leftarrow$ count_roles($G^{(m)}, \bar{\mathbf{R}}$)
end for
$\bar{\mathbf{C}} \leftarrow$ cosine_similarity($\bar{\mathbf{R}}$)

Algorithm 2 アンサンブルグラフ構築：ensemble_graphs($\mathcal{G}, S$)

Input: $\mathcal{G} = (G, p),\ G = (V, E),\ S$
Output: $\bar{G} = (V, \bar{E}, \bar{p})$
Initialize: $\forall e \in E,\ \bar{p}(e) \leftarrow 0$
Initialize: $\bar{E} \leftarrow \emptyset$
for $s = 1 : S$ **do** ▷ Sample and ensemble graphs
    $G_s = (V, E_s) \leftarrow$ sample_graph($\mathcal{G}$)
    for $e \in E_s$ **do**
        $\bar{p}(e)$ += $1/S$
    end for
    $\bar{E} \leftarrow \bar{E} \cup \{e\}$
end for

ある．$\delta(cond)$ は条件式 $cond$ が True の場合 1 を返し，False の場合 0 を返す Boolean 関数である．

つづいて，アンサンブルグラフ $\bar{G}$ に対して，Itzhack らのアルゴリズム [3] に基づいて連結 3 ノードを探索する（Algorithm 3）．Algorithm 3 において，$\Gamma(u) = \{v; (u, v) \in \bar{E} \wedge (v, u) \in \bar{E}\}$ はノード u の隣接ノード集合を表し，7 行目の $\bar{\Gamma}(u)$ は 6 行目の for 文で探索済みのノード集合を表す．すなわち，8 行目の $\Gamma(u) \setminus \bar{\Gamma}(u)$ は 6 行目で未探索なノード u の隣接ノード集合を表す．そして，探索した連結 3 ノード $G^{(m)}$ において，各ノードがどの Role として出現したかを，重み $\bar{p}(e)$ を考慮してカウントし，出現回数を要素とした Role ベクトル群（行列）$\bar{\mathbf{R}}$ を構築する（Algorithm 4）．具体的には，$G^{(m)}$ の 3 ノード u, v, w 間の 6 エッジの有無を 6 ビットのビット列 $\mathbf{b}_u$ で表す．ビット列からノード u がどの Role であるかを格納した辞書 Rcode に基づいて Role 番号を取得し（$i \leftarrow \mathrm{Rcode}(\mathbf{b}_u)$），ノード u の Role ベクトル $\bar{\mathbf{r}}$ の該当 Role 番号 i の値 $\bar{r}_{u,i}$ に加算する．加算する値は，6 エッジの有無と存在確率に基づき算出される $G^{(m)}$ の生起確率 $\Pr[G^{(m)}] = \prod_{e \in E^{(m)}} \bar{p}(e) \prod_{e \in E \setminus E^{(m)}} (1 - \bar{p}(e))$ である．また，有向 3 ノードの場合，ノード u に焦点を当てたビット列 $\mathbf{b}_u$ をビットシフトすることで，他の 2 ノード v, w に焦点を当てたビット列 $\mathbf{b}_v, \mathbf{b}_w$ を得ることができる．一方 4 ノード以上の場合は，単純なビットシフトではないが，同様にビット列を得ることができる．

そして，Role ベクトル間の類似度を要素とした行列 $\bar{\mathbf{C}}$ に基づいて，各ノードをクラスタに分類することで，$\bar{\mathbf{H}}$ を出力する．本手法では L 個のエッジを持った S 個のグラフのアンサンブルすることでアンサンブルグラフ $\bar{G}$ を求める．平均エッジ存在確率を p とすると，各サンプルグラフの期待エッジ数は pL となるため，$O(SpL)$ でアンサンブルグラフが得られる．アンサンブルグラフ 1 つに対して，Itzhack らのアルゴリズムにしたがって連結 k ノードを探索し，全 N ノードの Role 数をカウントする．したがって，k ノード Motif カウントの計算量と同じく，平均次数を $\bar{d}$ とすると，$O(N\bar{d}^{(k-1)})$) でアンサンブル・Role ベクトル $\bar{\mathbf{R}}$ が得られる．

Algorithm 3 連結 3 ノード探索：search_connected_triples(G)

Input: $G = (V, E, p)$
Output: Ψ
Initialize: $m \leftarrow 0$ ▷ Index for connected triples
Initialize: $\Psi \leftarrow$ list() ▷ Empty list
for $u \in V$ **do**
    for $v \in \Gamma(u)$ **do**
        $\bar{\Gamma}(u) \leftarrow \{v\}$
        for $w \in \big(\Gamma(u) \setminus \bar{\Gamma}(u)\big) \cup \Gamma(v)$ **do**
            $V^{(m)} \leftarrow \{u, v, w\}$
            $E^{(m)} \leftarrow V^{(m)} \times V^{(m)},\ \mathbf{p}^{(m)} = [p(e)]_{e \in E^{(m)}}$
            $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)})$
            $\Psi[m] \leftarrow G^{(m)}$
            m += 1
        end for
    end for
end for

Algorithm 4 Role カウント（更新）：count_roles($G^{(m)}, \bar{\mathbf{R}}$)

Input: $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)}), \bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leqq i \leqq 30}$
Output: $\bar{\mathbf{R}}$
$V^{(m)} \rightarrow \{u, v, w\}$
$\mathbf{b}_u \leftarrow$ motif2bits($G^{(m)}$)
$\mathbf{b}_v \leftarrow (\mathbf{b}_u \ll 2 | \mathbf{b}_u \gg 4) \& ((1 \ll 6) - 1)$
$\mathbf{b}_w \leftarrow (\mathbf{b}_v \ll 2 | \mathbf{b}_v \gg 4) \& ((1 \ll 6) - 1)$
$(i, j, k) \leftarrow (\mathrm{Rcode}[\mathbf{b}_u], \mathrm{Rcode}[\mathbf{b}_v], \mathrm{Rcode}[\mathbf{b}_w])$ ▷ Get role code
$\Pr[G^{(m)}] \leftarrow \prod_{e \in E^{(m)}} p(e) \prod_{e \in E \setminus E^{(m)}} (1 - p(e))$
$\bar{r}_{u,i}$ += $\Pr[G^{(m)}]$ ▷ Increment #roles
$\bar{r}_{v,j}$ += $\Pr[G^{(m)}]$
$\bar{r}_{w,k}$ += $\Pr[G^{(m)}]$

4.2 ベクトルアンサンブル法

Role ベクトルアンサンブル法（以下，VE 法）は，サンプルグラフ G_s から得られた Role ベクトル $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$ を平均することで，アンサンブル・Role ベクトル $\bar{\mathbf{R}}$ を生成する：

$$\mathcal{R} = \bigoplus_{G \in \mathcal{G}} (\mathbf{R}_G;\ \Pr[G]) \simeq \frac{1}{S} \sum_{s=1}^{S} \mathbf{R}_s = \bar{\mathbf{R}}.$$

そして，得られたアンサンブル・Role ベクトル $\bar{\mathbf{R}}$ からコサイン類似度 $\bar{\mathbf{C}}$ を計算し，類似度行列をもとに各ノードをクラスタに振り分け，$\bar{\mathbf{H}}$ を出力する．各サンプルグラフ G_s から Role ベクトル $\mathbf{R}_s$ を構築する際，最先端技術である LINC アルゴ

リズム [9] を用いる．LINC アルゴリズムでは，2 つのサンプルグラフ G_s と $G_{s'}$ におけるエッジ集合 E_s と $E_{s'}$ 間の差分 $D_{s,s'} = (E_s \setminus E_{s'}) \cup (E_{s'} \setminus E_s)$ に着目し，存在・不在の状態が変化したエッジ $e \in D_{s,s'}$ に関係する Role の出現数のみを更新する．平均エッジ出現確率を p とすると，状態変化するエッジ数の期待値は $2L(p-p^2)$ となるため，$p = 0.1$ や $p = 0.9$ のとき効果を発揮する．このようにして，S 個の $(N \times R)$ Role ベクトル $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$ を効率的に求め，平均を計算する．したがって，$\bar{m}$ を各エッジを含む平均 Motif インスタンス数とすると，VE 法により計算量 $O(S(L(p-p^2)\bar{m} + NR))$ でアンサンブル・Role ベクトル $\bar{\mathbf{R}}$ が得られる．

4.3 類似度アンサンブル法

類似度アンサンブル法（以下，SE 法）は，Role ベクトル群 $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$ から計算された類似度行列 $\{\mathbf{C}_1, \ldots, \mathbf{C}_S\}$ の平均を求め，アンサンブル類似度行列 $\bar{\mathbf{C}}$ を生成する：

$$\mathcal{C} = \bigoplus_{G \in \mathcal{G}} (\mathbf{C}_G;\ \Pr[G]) \simeq \frac{1}{S} \sum_{s=1}^{S} \mathbf{C}_s = \bar{\mathbf{C}}.$$

アンサンブル・類似度行列 $\bar{\mathbf{C}}$ に基づき各ノードをクラスタに振り分け，$\bar{\mathbf{H}}$ を出力する．ベクトルアンサンブル法と同様に，LINC アルゴリズムに基づき Role 数をカウントする．このようにして，S 個の $(N \times N)$ の類似度行列を求め，その平均を計算するため，本手法の主たる計算量は $O(SN^2)$ となる [1]．

4.4 クラスタアンサンブル法

クラスタアンサンブル法は，クラスタリング結果をアンサンブルし，所属クラスタ行列 $\{\mathbf{H}_1, \ldots, \mathbf{H}_S\}$ を生成する：

$$\mathcal{H} = \bigoplus_{G \in \mathcal{G}} (\mathbf{H}_G;\ \Pr[G]) \simeq \bigoplus_{s=1}^{S} (\mathbf{H}_s;\ 1/S) = \bar{\mathbf{H}}.$$

クラスタリング結果をアンサンブル手法はまだ確立されていないため，今回は触れない．

5 提案手法：拡張グラフアンサンブル法

拡張グラフアンサンブル法（以下，Ext-GE 法）では，確率的に Motif が崩壊し，他の Role に推移することを考慮して，各ノードの Role 頻度の期待値を計算する．図 4 に示す有向 3 ノードの Motif-Role の例にように，各 Role は含まれる Motif を構成するエッジ数により階層が定義できる．図 4 を見ると，6 エッジからなる Motif 13 および Role 30 が最上位にあり，2 エッジからなる Motif 1,2,4 および Role 1,2,4,5,10,13,16 が最下位にある．確率的にエッジの不在が起きることで，上位の Role は下位の Role へと変化し，下位 Role の出現頻度は増加する．したがって，S 個のサンプルグラフ G_s，$1 \leqq s \leqq S$ をアンサンブルした $\bar{G} = (V, \bar{E}, \bar{p})$ に対して，各ノードの Role 頻度をカウントする際，該当する Motif の下位 Motif を探索し，その Motif に含まれる Role 数も同時にカウントする（Algorithm 5）．

1：LINC アルゴリズムに基づき Role 数をカウントするが，類似度行列のアンサンブルにかかる計算量が支配的となる．

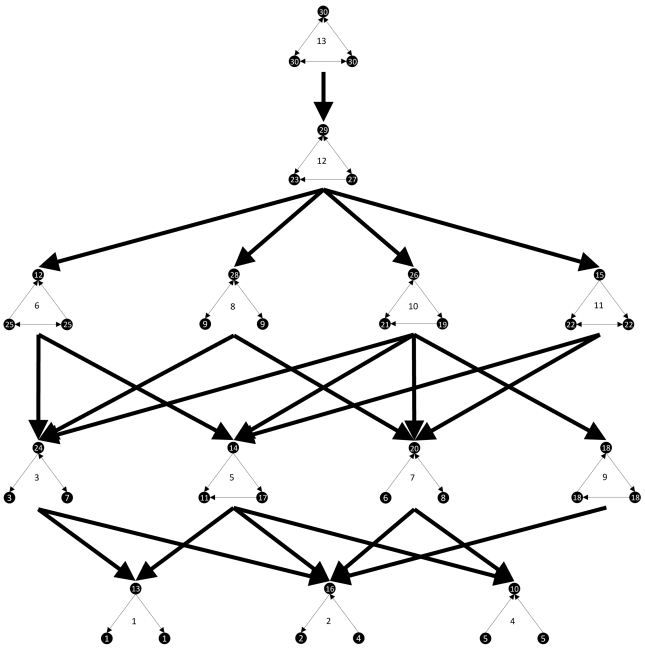


図 4　有向 3 ノード Motif における Role の推移

Algorithm 5 の 6 行目から 18 行目の While 文で，アンサンブルグラフにおいて探索された連結 3 ノードからなる Motif とその下位 Motif および Role を探索している．詳細には，4 行目の $\mathbf{b}$ がアンサンブルグラフにおいて探索された連結 3 ノード間のエッジ有無を表し，5 行目の 6 ビットすべてが 1 で初期化された $\mathbf{b}_u$ に対して，8 行目のビット AND 演算，17 行目の引き算を繰り返し行うことで，下位 Motif および Role を効率的かつ網羅的に探索している．この While 文は，6 ビットで表される有向 3 ノード Motif の場合，最大でも 64 反復する．その他の部分は，Algorithm 4 の count_roles 関数と同じである．また，11 行目の bits2motif 関数は，4 行目の motif2bits 関数の逆関数で，3 ノード間のエッジ有無を表すビット列から 3 ノード Motif の構造を獲得するためのものであり，擬似コードの便宜上載せている．Algorithm 6 に，Ext-GE 法の全体像を示す．GE 法（Algorithm 1）との違いは，8 行目の Role 数の期待値計算時に下位 Role への推移を考慮するか否かのみである．

6 評 価 実 験

6.1 データセット，実験設定

評価実験では，文献 [14] の手法によりロール数の真の期待値 $\mathcal{R}$，そこから計算した真の類似度行列 $\mathcal{C}$ を求める．そして，k-medoids クラスタリングの貪欲法により $\mathcal{H}$ を求め，真のクラスタリング結果とする．有効性評価として，著者らの既存手法である GE 法とともに，Uncertain グラフに対するモチーフカウントの最先端手法である LINC アルゴリズムを応用した VE 法，SE 法と本稿の提案法である Ext-GE 法に対して，各種法により計算した，アンサンブルロールベクトル $\bar{\mathbf{R}}$，アンサンブル類似度行列 $\bar{\mathbf{C}}$ に対して，真の値 $\mathcal{R}$，$\mathcal{C}$ との誤差 Root Mean Squared Error（以下，RMSE）を測定する．そして，アンサンブルクラスタリング結果 $\bar{\mathbf{H}}$ に対して，真のクラスタリング

Algorithm 5 拡張グラフアンサンブル法における下位 Role カウント（更新）：count_lower_roles($G^{(m)}, \bar{\mathbf{R}}$)

Input: $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)}), \bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leqq i \leqq 30}$
Output: $\bar{\mathbf{R}}$
$V^{(m)} \to \{u, v, w\}$
$\mathbf{b} \leftarrow$ motif2bits($G^{(m)}$)
$\mathbf{b}_u \leftarrow (1 \ll 6) - 1$
while True **do**
  if $\mathbf{b}_u < 0$ **then break end if**
  $\mathbf{b}_u$ &= $\mathbf{b}$
  $\mathbf{b}_v \leftarrow (\mathbf{b}_u \ll 2 | \mathbf{b}_u \gg 4) \& ((1 \ll 6) - 1)$
  $\mathbf{b}_w \leftarrow (\mathbf{b}_v \ll 2 | \mathbf{b}_v \gg 4) \& ((1 \ll 6) - 1)$
  $G^{(n)} \leftarrow$ bits2motif($\mathbf{b}_u, \mathbf{b}_v, \mathbf{b}_w$)
  $(i, j, k) \leftarrow$(Rcode[$\mathbf{b}_u$], Rcode[$\mathbf{b}_v$], Rcode[$\mathbf{b}_w$]) ▷ Get role code
  $\Pr[G^{(n)}] \leftarrow \prod_{e \in E^{(n)}} p(e) \prod_{e \in E \setminus E^{(n)}} (1 - p(e))$
  $\bar{r}_{u,i}$ += $\Pr[G^{(n)}]$ ▷ Increment #roles
  $\bar{r}_{v,j}$ += $\Pr[G^{(n)}]$
  $\bar{r}_{w,k}$ += $\Pr[G^{(n)}]$
  $\mathbf{b}_u$ −= 1
end while

Algorithm 6 拡張グラフアンサンブル法：Ext-GE($\mathcal{G}, S$)

Input: $\mathcal{G} = (G, p)$, $G = (V, E)$, S
Output: $\bar{\mathbf{C}}$
Initialize: $\bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leqq i \leqq 30} \leftarrow \mathbf{0}$
$\bar{G} \leftarrow$ emsemble_graphs($\mathcal{G}, S$)
$\Psi \leftarrow$ search_connected_triples($\bar{G}$)
for $G^{(m)} \in \Psi$ **do**
  $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)})$
  $\bar{\mathbf{R}} \leftarrow$ count_lower_roles($G^{(m)}, \bar{\mathbf{R}}$)
end for
$\bar{\mathbf{C}} \leftarrow$ cosine_similarity($\bar{\mathbf{R}}$)

$\mathcal{H}$ との類似性 Normalized Mutual Information（以下，NMI）を測定する．効率性評価として，各手法がアンサンブル類似度行列を出力するまでの実行時間を比較する．これらの実験により，厳密な期待値との誤差が少なく，実行時間が短い，本タスクに最も適した手法を調査する．

本実験で使用するグラフの基本統計量を表 1 に示す．Celegans は線虫の神経回路網，Gnutella は P2P ネットワーク，Blog はブログのトラックバックのネットワークである．これらのグラフの各エッジに対し，一様な出現確率 $p(e) = p \in \{0.1, 0.2, \ldots, 0.9\}$ を付与した．サンプル数は $S \in \{10^1, 10^2, 10^3, 10^4\}$ とし，クラスタ数は $K = 10$ の結果のみを示す [2] [3]．サンプル数やエッジの出現確率の違いにより，誤差や実行時間に影響があるか実験的に検証する．

表 1　使用する実有向グラフの基本統計量

データの種類	ノード数	エッジ数
Celegans [15]	131	764
Gnutella [15]	10,876	39,994
Blog	12,047	53,315

6.2 Role ベクトルの誤差評価

図 5 に，横軸をサンプル数 S，縦軸に RMSE を対数スケールでプロットしたものを示す．Celegans に対する結果を見ると，VE 法，Ext-GE 法ともにサンプル数に比例して誤差が少なくなる傾向にあることが確認できる．一方，GE 法に関しては誤差の値はサンプル数に依存せず一定である．また確率が $p = 0.1$ の際，他の確率と比較して誤差が 10^1 ほど小さくなっていることが確認できる．

Gnutella に対する結果を見ると，GE 法，VE 法，Ext-GE 法すべての結果が類似した結果となった．誤差の傾向は　GE 法を除いた残りの 2 つの手法は Celegans と同じで，サンプル数に比例して誤差が少なくなり，確率 $p = 0.1$ では他の結果より誤差が 10^1 ほど小さくなっていることが確認できる．

Blog に対する結果を見ると，Celegans と類似した結果であるが，GE 法とその他 2 つの手法との差が大きいことが確認できる．また，他の実験データより全体的に誤差が小さくなっていることも確認できる．

これらの結果から，VE 法，Ext-GE 法はサンプル数が多くなるほど誤差が小さくなる傾向にあり，GE 法は，サンプル数に依存せず常に一定の誤差があることがわかった．

6.3 類似度行列の誤差評価

図 6 に，横軸をサンプル数 S，縦軸を対数スケールで RMSE をプロットしたものを示す．Celegans に対する結果を見ると，VE 法，Ext-GE 法ともにサンプル数に比例して誤差が少なくなる傾向にあることが確認できる．一方，GE 法，SE 法に関してはサンプル数は誤差の値に依存せず一定な傾向にあり，確率が $p = 0.1$ の時にはサンプル数に反比例して誤差が大きくなっていることが確認できる．

Gnutella に対する結果を見ると，VE 法，SE 法，Ext-GE 法がサンプル数に比例して誤差が少なくなる傾向にある．一方，GE 法に関しては確率が $p = 0.1$ の時にはサンプル数に反比例して誤差が大きくなっている．

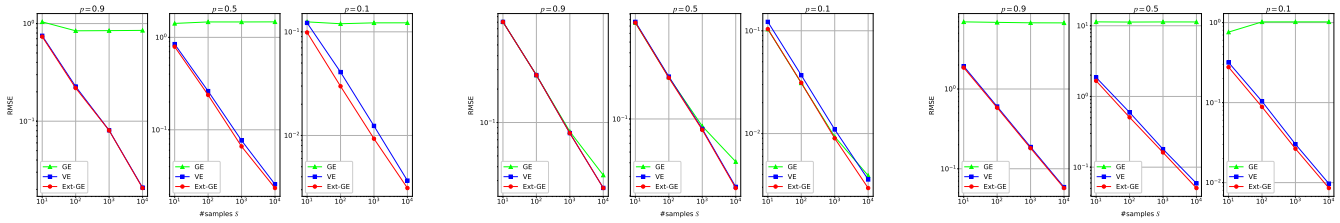
Blog に対する結果を見ると，Gnutella と同じく，VE 法，SE 法，Ext-GE 法がサンプル数に比例して誤差が少なくなる傾向にある．一方，GE 法に関しては確率が $p = 0.1$ の時にはサンプル数に反比例して誤差が大きくなっていることが確認できる．

これらの結果から，VE 法，Ext-GE 法はサンプル数が多くなるほど誤差が小さくなる傾向にあり，GE 法と SE 法は，サンプル数に依存せず一定の誤差があることがわかった．

6.4 クラスタリング結果の誤差評価

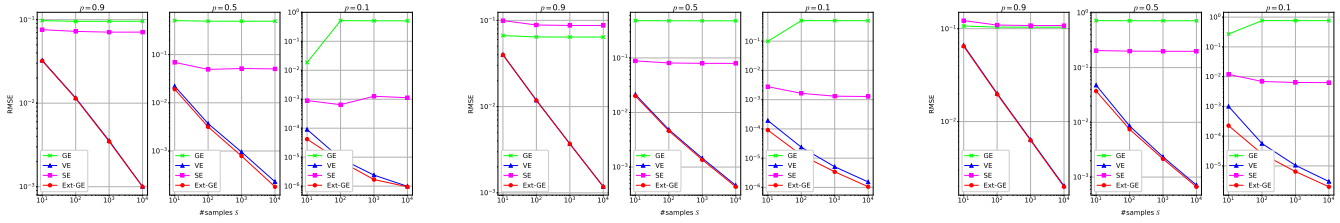
図 7 に，横軸をサンプル数 S，縦軸をクラスタリング結果の NMI をプロットしたものを示す．Celegans の結果を見ると，

2：他のクラスタ数における結果も同様であったため割愛する．
3：Celegans のみはグラフサイズが小さいため $K = 5$ とした．



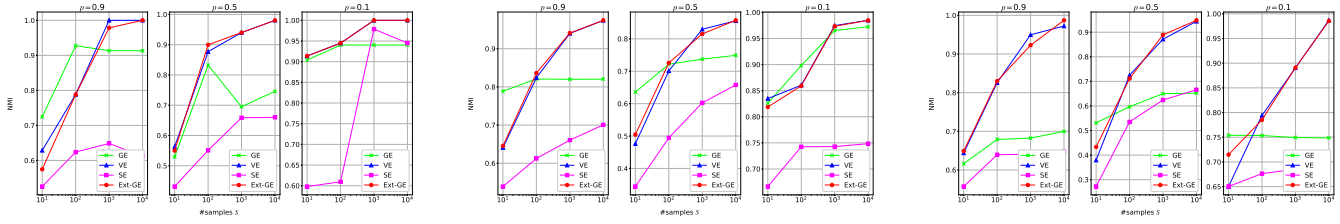
(a) Celegans (b) Gnutella (c) Blog

図 5　真のロールベクトル $\mathcal{R}$ と各種サンプリング法による近似値 $\bar{\mathbf{R}}$ の RMSE



(a) Celegans (b) Gnutella (c) Blog

図 6　真の類似度行列 $\mathcal{C}$ と各種サンプリング法による近似値 $\bar{\mathbf{C}}$ の RMSE



(a) Celegans (b) Gnutella (c) Blog

図 7　真のクラスタリング結果 $\mathcal{H}$ と各種サンプリング法による近似クラスタリング結果 $\bar{\mathbf{H}}$ の NMI

VE 法，Ext-GE 法がサンプル数に比例して類似度が高くなっており，サンプル数 $S = 10^4$ の時には，どの確率においても類似度が 1.00 に近い値となっていることが確認できる．GE 法は，サンプル数が少ないうちは全手法の中で最も類似度が高くなっているが，サンプル数 $S = 10^2$ を超えたタイミングで頭打ちとなり，類似度が下がってしまう傾向にあることがわかる．SE 法も GE 法と同様に，サンプル数が 10^3 を超えたタイミングで頭打ちとなり，類似度が下がる傾向が見られた．また，確率 $p = 0.1$ のサンプル数 $S = 10^2$ から 10^3 にかけて，類似度が急激に高くなっており，結果が安定していないことがわかる．Gnutella の結果を見ると，全手法とも，サンプル数に比例して類似度が高くなっていることが確認できる．特に，VE 法と Ext-GE 法はどの確率においても，安定して高い類似度を保持していることが確認できる．

Blog の結果を見ると，VE 法，Ext-GE 法は安定して，サンプル数に比例して高い類似度となっている．一方，GE 法，SE 法は他の実験データの結果と比較して類似度が小さくなっていることが確認できる．

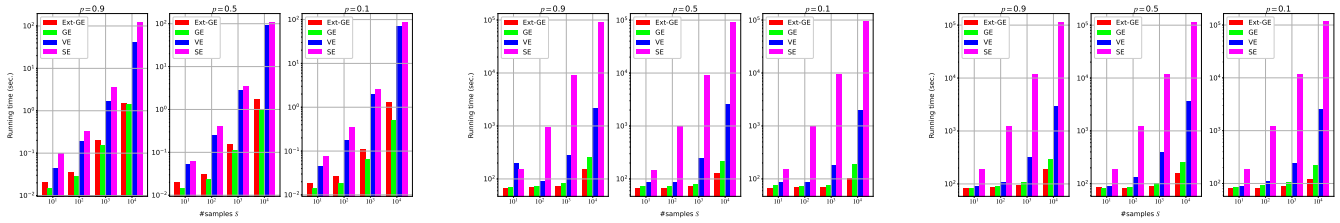
これらの結果から，VE 法，Ext-GE 法はサンプル数に比例して安定した類似度であり，一方，GE 法，SE 法はサンプル数，エッジの出現確率，グラフのサイズ等に大きく影響を受けることがわかった．

6.5　効率性評価

図 8 は，各種法がアンサンブル類似度行列 $\bar{\mathbf{C}}$ を出力するまでの実行時間を示す．横軸はサンプル数 S，縦軸は対数スケールで表した実行時間である．Celegans の結果を見ると，GE 法が最も実行時間が短いことが確認できる．Ext-GE 法がそれに次いで高速である．Gnutella の結果を見ると，Ext-GE 法が最も実行時間が短いことが確認できる．VE 法，SE 法はどちらもその他 2 手法と比較してかなり遅い結果となっていることが確認できる．Blog の結果を見ると，Gnutella と類似した結果であることが確認できる．傾向も同じであり，Ext-GE 法が他の手法より最も実行時間が短いことが確認できる．これらの結果から，GE 法と Ext-GE 法が最速な手法であり，データの種類やサンプル数，エッジの出現確率にも依存せず，安定していることがわかる．

7　おわりに

本稿では，既存タスクである Motif カウントを拡張した Motif-Role カウントを Uncertain グラフに対して行う問題を新たに定式化し，サンプリングしたグラフをアンサンブルする拡



(a) Celegans　　(b) Gnutella　　(c) Blog

図 8　各手法がアンサンブル類似度行列を出力するまでの実行時間

張グラフアンサンブル法を提案した．実験では，3 つの実ネットワークに一様なエッジ確率を付与したものを用い，既存手法と提案手法である拡張グラフアンサンブル法を比較して，厳密な期待値との誤差，実行時間の観点で評価した．結果として，本稿で提案した拡張グラフアンサンブル法が最先端手法である LINC を導入した既存手法であるベクトルアンサンブル法，類似度アンサンブル法や，それらより高速なグラフアンサンブル法と比較して，厳密な期待値に近い結果を出力し，かつ最も高速であることが確認できた．以上より，拡張グラフアンサンブル法が本稿で扱う問題に最も適していると結論づけた．

今後の課題としては，1) 有向 3 ノードモチーフ以外をもとにしたモチーフロールでのモチーフロール抽出，2) Hoeffding の不等式を用いた適切なサンプリング数の決定，3) クラスタリング結果のより詳細な分析，4) 各エッジに非一様な確率を付与した Uncertain グラフので実験，などがあげられる．

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．

文　献

[1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science (New York, N.Y.)*, Vol. 298, No. 5594, pp. 824–827, October 2002.

[2] S. Wernicke. A faster algorithm for detecting network motifs. In *Proceedings of the 5th International Conference on Algorithms in Bioinformatics*, WABI'05, pp. 165–177, Berlin, Heidelberg, 2005. Springer-Verlag.

[3] R. Itzhack, Y. Mogilevski, and Y. Louzoun. An optimal algorithm for counting network motifs. *Physica A: Statistical Mechanics and its Applications*, Vol. 381, pp. 482–490, 2007.

[4] J. A. Grochow and M. Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology*, RECOMB'07, pp. 92–106, Berlin, Heidelberg, 2007. Springer-Verlag.

[5] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield. Efficient graphlet counting for large networks. In *2015 IEEE International Conference on Data Mining*, pp. 1–10, 2015.

[6] A. Pinar, C. Seshadhri, and V. Vishal. Escape: Efficiently counting all 5-vertex subgraphs. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1431–1440, Republic and Canton of Geneva, CHE, 2017.

[7] 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.

[8] 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.

[9] 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.

[10] 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*, pp. 501–513, 2021.

[11] Malod-Dognin N. Yaveroğlu ö. N. Sarajlić, A. and N. Pržulj. Graphlet-based characterization of directed networks. *Nature*, Vol. 123, , 10 2016.

[12] Cheng-R. Huang Z. Fang Y. Hu, J. and S. Luo. On embedding uncertain graphs. In *ACM on Conference on Information and Knowledge Management*, Vol. 123, pp. 157–166.

[13] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An Analysis of Approximations for Maximizing Submodular Set Functions. *Mathematical Programming*, Vol. 14, pp. 265–294, 1978.

[14] A. Todor, A. Dobra, and T. Kahveci. Counting motifs in probabilistic biological networks. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, pp. 116–125, New York, NY, USA, 2015. Association for Computing Machinery.

[15] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.