複雑ネットワークにおける出現位置と役割に着目した効率的な成長誘発エッジ検出手法

稲福　和史† 　伏見　卓恭†† 　佐藤　哲司†††

† 筑波大学図書館情報メディア研究科　〒 305-8550 茨城県つくば市春日 1-2
††† 筑波大学図書館情報メディア研究系　〒 305-8550 茨城県つくば市春日 1-2
†† 東京工科大学コンピュータサイエンス学部　〒 305-0982 東京都八王子市片倉町 1404-1
E-mail: †{inafuku,satoh}@ce.slis.tsukuba.ac.jp, ††takayasu.fushimi@gmail.com

あらまし　現実の複雑ネットワークの多くは，日々エッジ（及びノード）の追加が行われその構造を変化させる動的ネットワークである．時間的に変化する動的ネットワークの変化点や変化・成長を誘発する構造を検出することは重要な課題である．本研究では，ユーザ間のコミュニケーションなどによって逐次的に追加されるエッジの出現時点におけるネットワーク構造を特徴量として，将来的な成長を誘発するエッジの抽出手法を提案する．提案手法では，ネットワークの成長を強化と拡張の 2 つの役割で捉え，エッジの出現位置と役割でモデル化する．前者は一般的にネットワーク全体の指標である近接中心性で定量化されるが，計算量が大きいことから局地的な指標として隣接スコアを新たに定義する．後者は，リンク元とリンク先の隣接ノード集合の類似度により，強化・拡張の役割を定量化する強化拡張スコアを定義する．いくつかの人工ネットワークと実際のデータを用いて提案する 2 種類のスコアを算出し，これらのスコアを用いてネットワークの成長を推定できることを検証した．その結果，情報拡散の性質をもつネットワークでは,，ネットワークの周縁部で強化する役割が高いエッジが成長を誘発する傾向にあることが明らかとなった．

キーワード　複雑ネットワーク，情報カスケード，Twitter

1 はじめに

ネットワークは，ノード (点) とそれらの繋がりを表すエッジ (線) を基本要素とするデータ構造である．現実世界の様々な事象や関係性はネットワークで表現できる．例えば，Twitter や Facebook などの SNS では，ユーザ同士のフォロー関係をフォローネットワークとして表せる．また，その他にも Web 上のハイパーリンクや道路網，商品の同時購入を表す共購買ネットワークなど様々な分野に渡ってネットワークが存在する．これらのネットワーク構造を理解することは，SNS のコミュニティ発見や交雑する道路の発見など，現実世界の課題を解決する有益な知見発掘に繋がる．そのため，無向ネットワークの指標として，平均クラスタ係数 [14] や平均ノード間距離 [17]，次数分布のべき指数 [1]，有向ネットワークの指標としてモチーフ [10] などが提案されるなど様々な研究が行われている．

その中でも，ノードやエッジの影響力を定量化するのは重要なタスクである．他のノードに対する影響力を定量化した既存指標として，期待影響度や媒介中心性 [6] などが挙げられる．期待影響度は自身が情報を発するときにどれだけのノードに届けることができるかを定量化したもの，媒介中心性はそのノードがネットワーク中の各 2 ノード間の最短経路上にどの程度出現するかを定量化したものである．しかし，期待影響度や媒介中心性をはじめとする指標の多くは，時間経過を考慮しない静的ネットワークを対象としたものである．一方，実世界の多くのネットワークは，時々刻々とノードとエッジが追加される動的ネットワークである．フォローネットワークなら新たなユーザや新しいフォロー関係が生まれる度に，共購買ネットワークなら新たな購買が行われる度に，ネットワーク構造は変化する．このような動的ネットワークについて，各種指標を算出する場合には大きく分けて 2 つ課題がある．

第一の課題は，そもそも実際の影響が発生する前に影響力を推定したいという点である．期待影響度や媒介中心性は，計算時点のネットワークに対して与える影響の指標である．しかし，現実での応用を考えるとノードやエッジが出現した時点でそれらが将来的にどの程度影響を与えるのかを推定したい．これが実現できれば，SNS マーケティングやコミュニティの成長予測などに大きく貢献する．第二の課題は，ネットワーク構造が変化する度に各指標の計算をやり直す必要があるため，計算コストが大きい点である．例えば，期待影響度は，伝達できるノード数の期待値を求めるため，全てのエッジについて行う切断・非切断の判定を充分な回数繰り返す必要がある．また，媒介中心性は，2 ノードの全組み合わせについて最短経路を求める必要があり，静的ネットワークですら大規模な場合には厳密解を求めることが困難である．このような静的ネットワークの指標を動的ネットワークに対してナイーブに適用し，構造変化の度に再計算を行うことは現実的でない．

これらの課題に対し，エッジ出現時の特徴を用いることで今後与える影響について分析できるのではないかと考えた．具体的には「ネットワークの何処に出現したか？」と「コミュニティを強化するか？拡張するか？」に着目する．とあるアイドルのファンからなるネットワークを想定する．このネットワー

クにおいては，中心部には密に結合する古参のファンが，周縁部には新参者のファンが存在していると考えられる．ここで，古参同士，新参同士など近い者同士の間で行われるコミュニケーションは，コミュニティの結合を強化する役割があるといえる．逆に，古参と新参の壁を超えて行われるコミュニケーションは，コミュニティを拡張する役割があるといえる．このように，ファン（ノード）間のコミュニケーションはエッジで表現される．また新たに生成されるエッジの役割によって，どの程度コミュニケーションを誘発できるか，その影響力は異なるのではないだろうか．

そこで，本研究では上述したエッジ出現時の 2 つの特徴を用いてネットワークに与える影響力の分析，高影響エッジの抽出に取り組む．ネットワークの性質により，成長を誘発するエッジの特徴は異なると考えられることから，人工データと複数の実データを用いて分析を行う．本研究の目的は，ネットワークの成長を誘発するエッジが出現時にどのような特徴を持つか明らかにし，それら高影響なエッジを抽出する手法を確立することである．

本論文の構成について説明する．まず 2 章で，ネットワーク上を情報が拡散する情報カスケードを扱った研究について説明し，本研究の立ち位置を明らかにする．その他，動的ネットワークの生成モデルや分析に関する研究について説明する．3 章で本研究で取り組む問題設定について確認する．4 章で本研究の提案手法について，ネットワーク上の位置を効率よく定量化する手法とネットワークの拡張・強化を定量化する手法の 2 つに大別して説明する．5 章で評価実験に用いるデータセットをまとめ，6 章で実験内容，結果について説明しそれらの考察を行う．最後に 7 章で本研究をまとめ，研究の貢献と課題について述べる．

2 関 連 研 究

動的ネットワークに関連する研究と本研究の立ち位置について説明する．人から人へと情報が伝わる時，情報は二者の間でやりとりされるだけでなく，受け手が新たな人へと伝達することでより広く拡散する．この情報伝達の連鎖現象を情報カスケード [2] と呼び，複雑ネットワーク上でこれをモデル化する研究が広く取り組まれている．以下，この情報カスケードに関連する研究について説明する．

Cheng ら [4] は，Facebook における写真共有において，カスケードがその後も成長を続けるかの予測を行っている．この研究では，時系列的特徴と構造的特徴がカスケードサイズ予測の重要な変数であることを示している．また，拡散する情報自体の分析として，川本ら [19] はマイクロブログ上での社会的影響力を持つ情報カスケードの早期検知を行っている．Twitter 上で広く拡散した社会的影響力（震災情報やデマ等）を持つツイートについて，テキスト特徴量やネットワーク特徴量が与える影響を明らかにしている．また，情報拡散モデルとして広く用いられる独立カスケードモデル [7] [8] [9] と線形閾値モデル [15] [16] について説明する．独立カスケードモデルは送信者中心型のモデルである．まず，エッジ毎に拡散確率を設定する．次に情報源となるノードから隣接するノードに対し，情報を伝達する．この時，情報伝達の成否は，はじめにエッジごとに設定した確率に独立に従うとする．これを繰り返すことで，情報拡散をシミュレートするのが独立カスケードモデルである．一方，線形閾値モデルは情報の受け取り手側のノードを中心に情報拡散をシミュレートする．各ノードには事前に重みの閾値が割り当てられる．情報源ノードから情報が伝達し，各ノードは受け取った重みが閾値を超えた場合にアクティブノードへと変化し，さらに隣接するノードへと情報を伝達する．吉川ら [18] は，上述した独立カスケードモデルや線形閾値モデルを拡張し，ソーシャルネットワーク上での期待影響度曲線を推定する手法を提案している．期待影響度とは情報が伝わったノード数を示す指標であり，これを事前に推定することで様々な応用が期待できる．情報拡散の系列データを EM アルゴリズムによって学習し，学習したモデルのパラメータを用いたシミュレーションによって期待影響度曲線を高精度に推定している．

特に Cheng ら [4] と吉川ら [18] の研究は，複雑ネットワークにおける情報拡散の規模を推定しようとする点において本研究と同じモチベーションを持つ．これらの研究との差異は，ネットワークの成長という観点に着目し，情報を届けたノード数ではなく誘発されたエッジ数を推定の対象とする点，エッジ出現時点で高速に算出できる特徴量を用いる点が挙げられる．

3 問題設定：誘発スコア

Twitter を始めとするソーシャルネットワークやメーリングリストネットワークでは，情報がノード間を連鎖的に伝播する情報カスケードが発生することが知られている．これは，ユーザをノード，インタラクション（Twitter のリプライやメールの送信など）をエッジとする動的ネットワークにおいては，エッジの出現が他のエッジの出現を誘発することであると捉えられる．本研究では，動的ネットワークで出現するエッジが，将来的にどの程度のエッジ出現を誘発するのかに着目する．情報拡散モデルの研究で用いられる線形閾値モデル [15] を応用し，**式 2** で示すエッジの**誘発スコア** $i_t(u,v)$ を定量化する．

図 1 を例に説明する．まず，**図 1(a)** のノード v_a, v_b, v_c とエッジ e_1, e_2 からなるネットワークを考える．ここでは，時刻 $t = 1$ において，ノード v_a から v_b に向けて，エッジ $e_1 = (v_a, v_b)$ が追加される．その後，同様に時刻 $t = 2$ において，エッジ $e_2 = (v_b, v_c)$ が追加されている．このネットワークでは，v_a から v_b，v_b から v_c へと連鎖的に情報が伝達されている．この時，$e_1 = (v_a, v_b)$ は $e_2 = (v_b, v_c)$ の出現を誘発しているといえる．すなわち，時刻 t に出現した v への入エッジが，時刻 t 以降のノード v からの出エッジを誘発するとみなす．このようにあるエッジが誘発した後続するエッジの数を基本的な誘発スコアとする．

出エッジが複数の入エッジによって誘発されるケースについて，**図 1(c)** を例に説明する．エッジ $e_2 = (v_b, v_d)$ に着目すると，時刻 $t = 3$ 以降に誘発される v_d からの出エッジは

$e_4 = (v_d, v_e)$, $e_5 = (v_d, v_f)$, $e_6 = (v_d, v_g)$,$e_7 = (v_d, v_h)$ の 4 本存在する．そのため，エッジ $e_2 = (v_b, v_d)$ の誘発スコアを 4 としたいところだが，v_d にはもう一本の入エッジ $e_3 = (v_z, v_d)$ が存在する．この時，v_d の 4 本の出エッジは，これらの 2 本の入エッジによって誘発されたと考え，誘発スコアを $e_2 = (v_b, v_d)$ と $e_3 = (v_z, v_d)$ の 2 本で分け合うこととする．

また，誘発したエッジが間接的に誘発したエッジも誘発スコアに含めることとする．例えば，上述したように $e_2 = (v_b, v_d)$ と $e_3 = (v_z, v_d)$ は v_d からの 4 本の出エッジを誘発している．そのうちの $e_5 = (v_d, v_f)$ に着目すると，その先に更に $e_9 = (v_f, v_i)$ と $e_{10} = (v_i, v_e)$ の 2 本のエッジが存在することが分かる．これらもまた，$e_2 = (v_b, v_d)$ と $e_3 = (v_z, v_d)$ とによって誘発されたエッジであるとみなせる．最終的に，$e_2 = (v_b, v_d)$ と $e_3 = (v_z, v_d)$ は合計で 6 本のエッジを誘発しているといえる．さらに 2 本のエッジでそれらを分け合うため，誘発スコアはそれぞれ 3.0 となる．

$$s_t((u,v)) = \frac{|\,\mathrm{OV}(v)|}{|\,\mathrm{IV}(v)|} \tag{1}$$

$$i_t((u,v)) = \begin{cases} 0 & (|\,\mathrm{OV}(v)| = 0) \\ s_t(u,v) + \frac{(\Sigma_{x \in \mathrm{OV}(v)}\{s_t((v,x)) + i_t((v,x))\})}{|\,\mathrm{IV}(v)|} & (|\,\mathrm{OV}(v)| > 0) \end{cases} \tag{2}$$

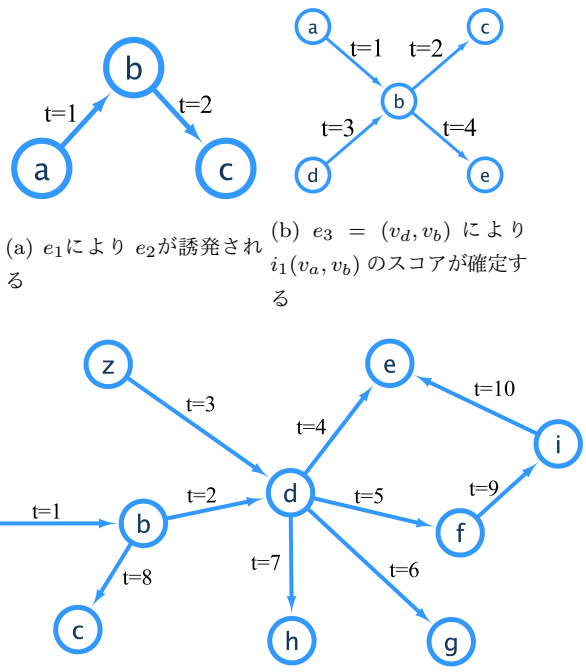
また，**図 1(b)** のようなケースにおいて上述した算出方法では，$e_1 = (v_a, v_b)$ の誘発スコアが 2 となるように古いエッジほど誘発スコアが次々と累積され大きくなってしまう問題がある．これに対処するため，各エッジについてスコアを確定するタイミングを設ける．ノード v への入エッジの誘発スコアは，スコアが 0 を上回っている状態で，新たにノード v への入エッジが出現したタイミングで確定する．

図 1(b) で時系列に構造変化を追いながら説明する．まず，時刻 $t = 1$ で $e_1 = (v_a, v_b)$ が出現する．この時点で誘発スコア $i_1(v_a, v_b)$ は 0 である．次に，時刻 $t = 2$ で $e_2 = (v_b, v_c)$ が出現する．これは誘発されて出現されたエッジであるので誘発スコア $i_1(v_a, v_b) = 1$ となる．時刻 $t = 3$ で v_b に対して新たな入エッジ $e_3(v_d, v_b)$ が発生する．この時 $e_1(v_a, v_b)$ の誘発スコアは 1 なので，スコアが確定する．時刻 $t = 4$ で新たなに $e_4 = (v_b, v_e)$ が出現するが，これは $e_3 = (v_d, v_b)$ のみによって誘発されたエッジであるとみなす．これにより，古いエッジほど誘発スコアが大きくなることを防ぐ．

誘発スコアはネットワークの成長と共に累積していく値であるが，本研究ではこれをエッジ出現時の特徴から分析・推定することを目指す．

4 提 案 手 法

1 節でも説明したように，動的ネットワークにおけるノードやエッジの影響力を定量化・予測することは重要なタスクである．我々は，エッジがどの位置に出現したのか，そしてネットワーク（コミュニティ）を強化したのか拡張したのかという特徴が，ネットワークの成長と大きく関わると考えた．そこで，エッジの「出現位置」と「強化・拡張」を定量化する手法を提案する．まず，本研究で用いる用語や変数について整理する．その後，ネットワーク上の「位置」を効率よく定量化する手法，リンク元とリンク先の関係に基づき「強化・拡張」を定量化する手法について説明する．



(a) e_1により e_2が誘発される　(b) $e_3 = (v_d, v_b)$ により $i_1(v_a, v_b)$ のスコアが確定する

(c) 複数のエッジによって誘発される場合及び誘発が連鎖する場合

図 1: 誘発スコアの模式図

4.1 準　　備

本研究で用いる用語や変数について整理する．提案手法は，ノード集合 V と有向エッジ集合 E からなる有向グラフ $G = (V, E)$ を対象とする．ここで，ノードは SNS などにおけるユーザを表し，有向エッジ $e = (u, v)$ はユーザ u から v へ情報を伝達するなどのコミュニケーションを表す．特に，本稿では時間とともにエッジが 1 本ずつ追加されることで成長する動的グラフを対象とする．時刻 t までに行われたコミュニケーションを表す有向エッジの集合を $E^{(t)} = \{e_1, e_2, ..., e_t\}$ で表す．このとき，はじめてコミュニケーションを行ったノードを追加したノード集合 $V^{(t)}$ を定義する．そして，グラフ $G^{(t)} = (V^{(t)}, E^{(t)})$ におけるノード u の隣接ノード集合を $NV^{(t)}(v) = \{(u,v) \in E^{(t)} \vee (v,u) \in E^{(t)}\}$ と表記する．

4.2 ネットワーク上の位置の推定手法：隣接スコア

ネットワーク上の位置を求める方法として，近接中心性 (Closeness Centrality) [6] を用いることが考えられる．しかし，近接中心性は計算量が大きく，ネットワークを更新する度に最

初から計算をやり直すのは現実的ではない．そこで，新規のエッジがネットワーク上のどの位置に出現したかを高速に近似する手法を提案する．

近接中心性は多くのノードにより短い距離で到達できるほど中心部だとする指標である．極端なケースとして，自ノードからすべての他ノードに直接リンクしている場合，近接中心性は 1 になる．すなわち，直リンク数が多いほど中心部に位置しやすいといえる．そこで，隣接ノード集合のサイズが大きいほど中心部に位置するとみなす**隣接スコア**を提案する．本研究では，エッジ毎にスコアを算出したいので，エッジの両端から隣接ノード集合を取得する．それらの和集合サイズをネットワーク全体のノード数で除して正規化した値を隣接スコアとし，ネットワーク上での位置指標として用いる．具体的には，有向エッジ $e_t = (u, v)$ の隣接スコア $n_t(u, v)$ を時刻 t におけるノード u と v の隣接ノード集合を用いて**式 3** のように定義する．

$$n_t(u, v) = \frac{|NV^{(t)}(u) \cup NV^{(t)}(v)|}{|V^{(t)}|} \tag{3}$$

また，隣接ノード集合の大きさはノードの次数に等しい．そのため，本手法は次数中心性 [6] の拡張であるともいえる．

4.3 ネットワークの強化と拡張：強化拡張スコア

情報拡散ネットワークにおいて新たなエッジが出現するとき，リンク元とリンク先との関係によりそのエッジの役割は大きく異なる．例えば，リンク元とリンク先が同一のコミュニティを形成している場合には，エッジの役割はおしゃべりなどのコミュニティを強化する情報伝達である．逆に，リンク元とリンク先の関係が薄い場合には，情報拡散やはじめましての挨拶などのコミュニティを拡大する役割だといえる．本節では，このようなエッジの役割を定量化する**強化拡張スコア** $j((u, v))$ を提案する（**式 4**）．

具体的には，エッジ両端のノードに関する隣接ノード集合の Jaccard 係数を求めることで，リンク元とリンク先の関係の強さを定量化する．

$$j((u, v)) = \frac{|NV(u)) \cap NV(v)|}{|NV(u) \cup NV(v)|} \tag{4}$$

Jaccard 係数は 2 つの集合の類似度を測る指標であり，共通集合を和集合で除することで算出する．あるエッジについて，2 つの隣接ノード集合の重複が多いほど関係が強く，逆に重複が少ないほど関係の薄いノード間をつなぐエッジとなる．すなわち，エッジ $e_t = (u, v)$ について，Jaccard 係数 $j(e_t)$ が 1 に近いほどネットワークの強化，0 に近いほどネットワークの拡張を意味する．

5 データセット

本研究の評価実験では，Connecting Nearest Neighbor モデルで生成した人工ネットワークと Twitter のインタラクションデータセット及び電子メールのやりとりのデータセットを用いる．ここでは，これらのデータセットについて説明する．

ネットワークのデータソース，呼称，規模について**表 1** に示す．実データにおいては，同一のノードペア間に同じエッジが付与される重複エッジが存在する．これを異なるエッジとみなしエッジ数をカウントしたものが動的エッジ数，同じエッジであるとみなすのが静的エッジ数である．それぞれのネットワークの詳細について，以下で説明する．

表 1: ネットワークのノード数及びエッジ数

ネットワーク名	ノード数	静的エッジ数	動的エッジ数
CNN1K-NW	497	994	994
CNN10K-NW	5,032	9,994	9,994
Reply-NW	1,233	1,622	1,971
Mention-NW	31,947	44,566	51,472
Retweet-NW	45,804	62,817	74,380
Eucore-NW	984	24,926	332,330

5.1 Connecting Nearest Neighbor (CNN) モデル

Connecting Nearest Neighbor モデル (以下，CNN) は，Vazquez [13] により提案された「友達の友達は友達になる」性質を有するネットワーク生成モデルである．新たなエッジとノードを追加する処理，ポテンシャルリンクを実リンクへと変換する処理を確率 p で選択することを繰り返しネットワークを構築する．ポテンシャルリンクとはいわばエッジ候補のことであり，これらからエッジを選ぶことで「友達の友達は友達になる」性質が生まれている．処理の選択を行う確率パラメータを $p = 0.5$，ループ回数 l を 100, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 100,000 とし，各ループ回数毎にそれぞれ 10 種，合計 120 個のネットワークを生成した．ネットワークの規模はループ回数 l に比例して大きくなる．また，$p = 0.5$ であるので，エッジ数はおよそループ数と等しく，ノード数はループ数の $\frac{1}{2}$ 程度である．加えて CNN モデルでは重複エッジが発生しないので，静的エッジ数と動的エッジ数の数は等しい．生成した 120 個のネットワークのうち，**表 1** には，ループ数 l=1000 の CNN1K-NW とループ数 l=10,000 の CNN10K-NW を示した．

5.2 Higgs Twitter Dataset

Higgs Twitter Dataset[1] [5] は，2012 年 7 月にヒッグス粒子が発見された際の，ヒッグス粒子に関するツイートのインタラクション（リプライ，メンション，リツイート）を収集したデータセットである．具体的には，2012 年 7 月 1 日から 2012 年 7 月 7 日の期間における，lhc, cern, boson, higgs のキーワードを含むリプライツイートとメンションツイートの送信ユーザと宛先ユーザ，キーワードを含むツイートの投稿者とそれをリツイートしたユーザ，各インタラクションの発生時刻が記録されている．つまり，時刻 t にユーザ u がユーザ v へとリプライを送ると，ユーザ u からユーザ v に有向エッジ $e_t = (u, v)$ が付与される．このようなエッジの付与をインタラクションの発生順に行うことで，動的なネットワークを構築する．本研究では，

1 : https://snap.stanford.edu/data/higgs-twitter.html

インタラクション種別に 3 種のネットワークを構築した．各ネットワークの規模を**表 1** に示す．なお，本研究では構築されたネットワークのうち，最終時刻における最大連結成分を抽出している．また，本節冒頭でも言及したとおり，実データには同じノードペア間に繰り返しエッジが付与される重複エッジが存在するが，本研究ではこれらは異なるエッジとして扱う．リプライツイートからなるネットワークを Reply-NW，メンションツイートからなるネットワークを Mention-NW，リツイートからなるネットワークを Retweet-NW とする．

5.3 email-Eu-core temporal network

email-Eu-core temporal network[2] [11] は，ヨーロッパの研究機関における電子メールのやり取りを収集したデータセットである．時刻 t にユーザ u からユーザ v にメールを送信したとき，有向エッジ $e_t = (u, v)$ が記録されている．データの収集期間は 803 日間であり，研究機関外との送受信は含まれていない．本データセットから構築したネットワークを Eucore-NW とする．こちらも 5.2 節と同様，最終時刻における最大連結成分を抽出しており，重複エッジは異なるエッジとして扱っている．

6 評価実験

本節では 5 節で構築したネットワークを用いて，提案手法の有効性を評価する．隣接スコアと強化拡張スコア，誘発スコアの関係について示す．

6.1 隣接スコアはネットワーク上の位置を効率よく定量化できるか

4.2 節で，エッジ両端の隣接ノード集合からネットワーク上の位置を効率よく定量化する隣接スコアを提案した．この手法について，本節では効率性と正確さの観点から評価を行う．

また 5 節で説明したように，CNN は規模別に 12 種類，さらにそれぞれの規模について 10 パターンの異なるネットワークを生成している．すなわち，CNN によって生成されたネットワークは合計で 120 個存在する．本節においては，各ネットワークに対し様々な評価指標を算出するが，特に断りの無い限り，CNN の評価指標は各規模毎に平均値を取るものとする．

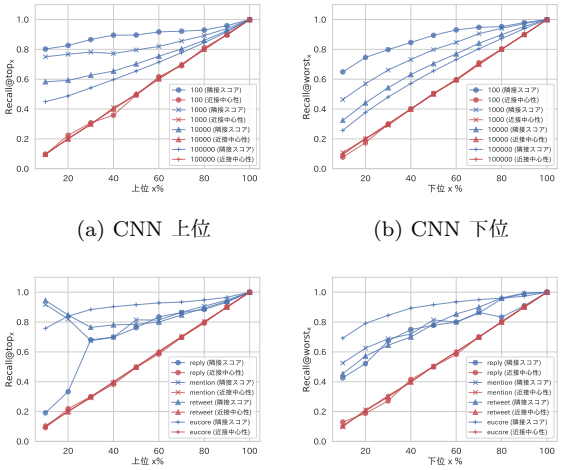
6.1.1 エッジの調和近接中心性

隣接スコアが正しくネットワーク上の位置を定量化しているか評価するために，近接中心性との比較を行う．近接中心性は，あるノードから他のノードへ平均的にどれくらい近いか示す指標である．しかし，最もよく知られている Freeman による近接中心性 [6] は，ネットワークが一つの連結成分から構成される必要があることから，比較に用いることができない．そこで，非連結なノードとの最短経路長は無限であるとみなし，調和平均を用いてスコアを算出する調和近接中心性 [12] [3] を用いる．また，隣接スコアはエッジに関する指標であるのに対し，調和近接中心性はノードに対して求められるスコアである．そのため本研究では，エッジ両端ノードの調和近接中心性スコアの平均値を用いる．

6.1.2 近似の妥当性

5 節で構築したネットワークを対象に隣接スコアと近接中心性を算出し，その妥当性を評価する．具体的な評価指標として，上位 (下位)x%のエッジを抽出した際の再現率を算出した．近接中心性上位 x%(及び下位 x%) のエッジを正解エッジとし，隣接スコア上位 x%(及び下位 x%) のエッジを抽出する時，正解エッジをどの程度抽出できるか（再現率）を検証する．CNN4 種 (ループ回数 $l = 100, 1,000, 10,000, 100,000$)，Reply-NW, Mention-NW, Retweet-NW, Eucore-NW，合わせて 7 つのネットワークで実験した．また，ベースラインとしてランダムにエッジを抽出した際の再現率も併せて算出した．**図 2** に正解エッジ割合に対する再現率の推移を示す．横軸が正解エッジの割合，縦軸が再現率，青いプロット線が提案手法に，赤い線がランダム抽出に対応する．

上位 10%のエッジを正解とすると，ランダム抽出したエッジに正解が含まれる確率は 10%である．赤いプロット線に着目すると，実際に再現率はおよそ x%を示し，一定のペースで推移している．一方，提案手法について，CNN ではネットワークの規模が大きいほど再現率が下がる傾向にあるものの，ランダム抽出のそれを大きく上回っている．また，実データも全てのネットワークでベースラインを上回る結果となった．特に**図 2(c)**, **図 2(d)** の Eucore-NW は，上位・下位ともに 8 割近い再現率を示している．また，Reply-NW, Mention-NW, Retweet-NW もベースラインと比較して概ね 20%程度高い再現率を示している．Reply-NW のみ，$Recall@worst_{10}$，$Recall@worst_{20}$ が低めであるものの，$Recall@worst_{30}$ 以降は大きくスコアを伸ばしている．これらのことから，近接中心性の上位（下位）エッジ抽出について，提案手法の隣接スコアは有効だといえる．



(a) CNN 上位　(b) CNN 下位

(c) Higgs 及び Eucore 上位　(d) Higgs 及び Eucore 下位

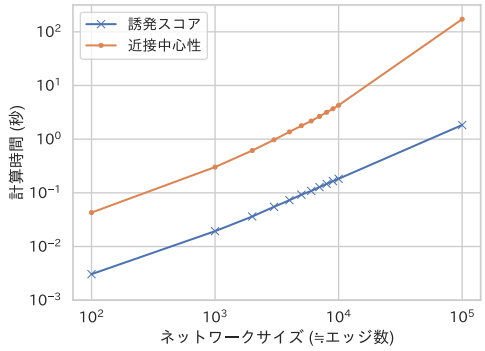
図 2: 上位/下位 x%抽出時の再現率

6.1.3 計算効率

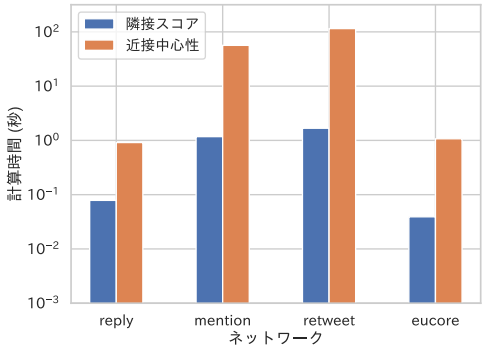
5 節で構築したネットワークを対象に隣接スコアと近接中心性を算出し，その処理時間を比較する．CNN 全 12 種と Reply-NW, Mention-NW, Retweet-NW，Eucore-NW について，両

2 : https://snap.stanford.edu/data/email-Eu-core-temporal.html

指標の計算時間を**図 3** に示す．CNN についてはネットワーク数が多いことから，折れ線グラフで示した．横軸がネットワークの規模及び種別，縦軸が計算にかかった秒数を示し，青い線（バー）が隣接スコア，オレンジの線（バー）が近接中心性に対応する．まず，CNN について，どの規模においても隣接スコアが 10 倍から 100 倍程度高速に動作していることが分かる．また規模が大きくなるほど，処理時間の差も大きくなる傾向にある．実データについても，比較的小規模な Reply-NW の処理時間差は 10 倍程度，最も大規模な Retweet-NW では 100 倍近い差となっており，CNN と同様である．このことから，提案手法は近接中心性よりも効率よくネットワーク上の位置を定量化できるといえる．



(a) CNN



(b) Higgs 及び Eucore

図 3: 隣接スコアと近接中心性の計算時間比較

6.2 誘発スコアの推定

本節では隣接スコアと強化拡張スコアを用いて，後続するエッジの誘発スコアについて分析する．分析対象は，CNN1K-NW, CNN10K-NW，Reply-NW, Mention-NW, Retweet-NW，Eucore-NW とした．なお，CNN は各規模毎に 10 パターン存在するが，本節では**表 1** に示したネットワークを用いる．

また，本研究で用いるネットワークは，時系列のエッジリスト $E = \{e_1, e_2, ..., e_T\}$($T$ は最終時刻) によって与えられる．本節では，このエッジリストのうち前半の 25%と後半の 25%を除いた中央部分 50%のエッジについてのみ分析に用いる．これは，エッジの出現時期による不均衡を補正するための処理である．例えば，最初に出現する e_1 の隣接スコアは，ネットワーク G 中に e_1 を構成する 2 ノードしか存在しないため，必ず 1 になる．また，最終時刻 T に出現するエッジ e_T は，その後に発生するエッジがデータセット中に存在しないため，誘発スコアが 0 以上になることはない．このようにデータセットが有限であるため生じる問題を回避するため，データセットの中央部分を分析に用いることとした．なお，**表 1** に示すネットワークの規模や，誘発スコアは最終時刻 T 時点のものである．

表 2: 重回帰分析の決定係数・偏回帰係数

ネットワーク	決定係数 R^2	偏回帰係数	
		隣接スコア	強化拡張スコア
CNN1K-NW	0.167	1.051	0.820
CNN10K-NW	0.170	1.080	0.879
Reply-NW	0.154	-0.337	0.418
Mention-NW	0.654	-0.680	0.520
Retweet-NW	0.699	-0.761	0.455
Eucore-NW	0.042	0.660	-0.249

6.3 隣接スコア，強化拡張スコア，誘発スコアの関係

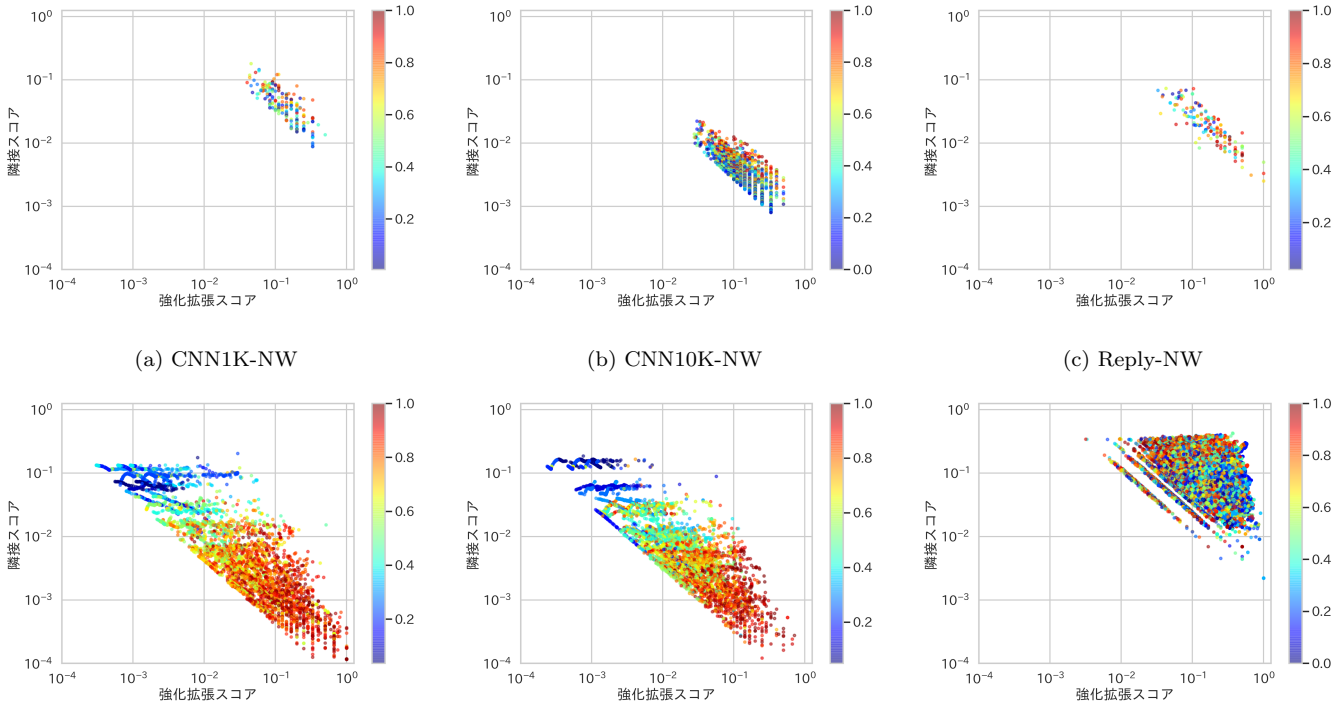
隣接スコア，強化拡張スコア，誘発スコアの関係を分析する．

3 指標の散布図を**図 4** に示す．横軸が強化拡張スコア，縦軸が隣接スコア，プロット点の色が誘発スコアの色に対応している．なお，誘発スコアはランキングをとった上で 0-1 の範囲に収まるように正規化を行った．色が赤いほど誘発スコアが高く，青いほど誘発スコアは低い．また，誘発スコア $i(e) = 0$ のエッジについてはノイズとして除外している．

図 4(d) の Mention-NW と**図 4(e)** の Retweet-NW をみると，左上から右下にかけて色が青から赤へと顕著に変化している．これはネットワークを周縁部で強化するエッジほど，誘発スコアが高くなる性質を持つことを意味する．これらのエッジは，いわばネットワークの周縁部で新たなコミュニティ形成を担うエッジだと考えられる．コミュニティそのものを誘発することから，誘発スコアが大きくなる傾向になるといえる．

この結果をより定量的に評価すべく，隣接スコアと強化拡張スコアを説明変数，誘発スコアを目的変数として重回帰分析を行った．この際，各変数のべき乗分布を考慮しいずれも常用対数をとった．自由度調整済み決定係数 R^2 と各説明変数の偏回帰係数を**表 2** に示す．Mention-NW と Retweet-NW の決定係数はいずれも 0.7 弱であり，一般に説明力が高いといえる数値である．ここで重要なのは，説明変数の隣接スコアと強化拡張スコアはいずれもエッジの出現時点で算出される指標であるのに対し，目的変数の誘発スコアは最終時刻に算出される指標である点である．すなわち，Mention-NW と Retweet-NW に関しては，エッジの出現時点で大きな誘発スコアを持つエッジの推定が可能だといえる．

一方，CNN10K-NW，CNN-100K-NW，Reply-NW，Eucore-NW については目論んだ結果は得られなかった．この理由は 6.4 節で考察する．



(a) CNN1K-NW　(b) CNN10K-NW　(c) Reply-NW

(d) Mention-NW　(e) Retweet-NW　(f) Eucore-NW

図 4: 隣接スコア，強化拡張スコア，誘発スコアの関係 ($(i(e) > 0)$)

6.4 考　　察

本節では，主に 6.2 節で行った実験について考察を行う．

まず，**図 4** で顕著に傾向が現れ重回帰モデルが高い精度を示した Mention-NW と Retweet-NW について考察する．これらのネットワークで多くのエッジを誘発する傾向にあったのは，ネットワークを周縁部で強化するエッジであるが，これらは具体的にどのような役割を果たしたのだろうか．そもそも，多くの影響を与えることは，多くの会話や情報発信を誘発することを意味する．ネットワークの周縁部ではノードやエッジが少ないことから，中心部と比べて自身が他者に与える影響は大きい．その上で親しい相手とリンクしネットワークを強化したということから，これらのエッジは周縁部の小規模なコミュニティの起点であったと考えられる．コミュニティの起点は，その後のコミュニティ全体を誘発するので，必然的に大きな誘発スコアを示す．また，メンションやリツイートはシェアの意味合いが強い行動でもある．すなわち，情報が狭いコミュニティに閉じず拡散される性質も併せ持つ．これらを踏まえると，Mention-NW や Reply-NW は，ネットワークの周縁部における小規模なコミュニティの出現・成長を促し，さらにそのコミュニティの周縁部に新たなコミュニティが発生することで，成長を繰り返すネットワークだと考えられる．そのため，周縁部におけるコミュニティ形成に貢献したエッジが高い誘発スコアを得たと考えられる．

一方，CNN1K-NW，CNN10K-NW，Reply-NW，Eucore-NW では，提案手法と誘発スコアの間に明確な関係は見られなかった．この原因について考察する．まず，Reply-NW や Eucore-NW は，メンションやリツイートと異なり情報の拡散が発生しにくいネットワークだと考えられる．例えば，Eucore-NW は電子メールのネットワークであるが，拡散する必要のあるメールは (データセットの対象外である) メーリングリストなどで一斉に配信されるのではないか．また，Twitter におけるリプライは，そもそも限られたユーザ同士のやり取りのための機能だといえる（多くのユーザと共有するためにはリツイートやメンションが用いられる）．すなわち，Reply-NW や Eucore-NW は，本研究で議論するコミュニケーションが拡散していくような変化に乏しいネットワークである．そのため，誘発スコアによって成長を表せなかったと考えられる．また，CNN モデルのネットワークについて，ノード追加時に設定されるエッジがポテンシャルリンクを誘発すると想定していたが，期待通りの挙動を示さなかった．これについては，ノード追加処理と実リンク変換処理の確率パラメータ p の設定により好転する可能性がある．後者の処理が多くなるように p を設定することで，誘発されるエッジが増え，提案手法と誘発スコアの間に何らかの関連が観察できると期待している．また，異なるアプローチとして CNN で設定したネットワークを用いた，情報拡散のシミュレートが挙げられる．より情報拡散の文脈に近い人工データを用いることで，提案手法の有効性や改善など様々な示唆が得られると考えている．

7 ま と め

現実世界の複雑ネットワークは，時々刻々と変化する動的ネットワークである．これらのネットワークにおいて，後続す

るノードやエッジを誘発する高影響な構造を推定・抽出することは重要なタスクである．本研究では，出現時の特徴を用いてネットワーク成長を誘発するエッジを効率的に抽出する手法を提案した．具体的には「エッジの出現位置（隣接スコア）」と「ネットワークを強化したか，拡張したか（強化拡張スコア）」の 2 点を用いた．Twitter の情報拡散ネットワークを用いた実験により，ネットワークを周縁部で強化するエッジがその後多くのエッジを誘発することを明らかにした．また，提案手法が既存手法に比べ高速に動作することを示した．

今後の課題は次の通りである．まず，情報拡散をシミュレートした人工データでの検証が挙げられる．情報拡散の文脈に近い人工ネットワークを用いることで，提案手法の改善に大きな示唆が得られると考えている．また，本研究ではネットワーク構造のみを特徴量とする手法を提案した．しかし，実際の情報拡散を分析する上では，情報の内容そのものに着目することも重要である．ネットワークベースの手法とコンテンツベースの手法を相補的に用いることで，より精緻な分析や推定モデルの構築が可能だと考えている．

謝　　辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです．ここに記して謝意を示します．

文　　献

[1] A. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys*, pages 47–97, 2002.

[2] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992.

[3] P. Boldi and S. Vigna. Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262, 2014.

[4] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. ACM, 2014.

[5] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3:2980, 2013.

[6] L. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[7] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[8] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[9] M. Kimura, K. Saito, and H. Motoda. Blocking links to minimize contamination spread in a social network. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(2):9, 2009.

[10] 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.)*, 298(5594):824–827, 2002.

[11] A. Paranjape, A. R. Benson, and J. Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 601–610, New York, NY, USA, 2017. ACM.

[12] Y. Rochat. Closeness centrality extended to unconnected graphs: The harmonic centrality index. Technical report, 2009.

[13] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104+, 2003.

[14] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[15] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[16] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

[17] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[18] 吉川友也, 斉藤和巳, 元田浩, 大原剛三, and 木村昌弘. 情報拡散モデルに基づくソーシャルネットワーク上でのノードの期待影響度曲線推定法. **電子情報通信学会論文誌** *D*, 94(11):1899–1908, 2011.

[19] 川本貴史, 豊田正史, and 吉永直樹. マイクロブログにおける社会的影響力を持つ情報カスケードの早期検知に向けて. **第** *8* **回** *Web* **とデータベースに関するフォーラム論文集**, pages 48–55, 2015.