図書館情報メディア研究科修士論文

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抽出に関する研究

2020年3月

201821606

稲福 和史

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抽出に関する研究

筑波大学
図書館情報メディア研究科
2020年3月
稲福 和史

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抽出に関する研究

A Study on Efficient Extraction of Growth-Inducing Edges Based on Their Appearance Positions and Roles in Complex Networks

学籍番号：201821606
氏名：稲福 和史
Inafuku Kazufumi

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．これらのネットワークを解析し，ネットワークの成長を誘発する，すなわちネットワークをより成長させる構造を抽出することは重要な課題である．本研究では，エッジ出現時点の特徴量を用いて，将来的にネットワークの成長を誘発するエッジの抽出手法を提案した．具体的には，「エッジの出現位置」と「ネットワークを強化したか，拡張したか」の 2 つの観点に着目する．前者について，一般にネットワーク上の位置指標としては近接中心性が用いられるが，これは計算量が大きく，常に変化する現実の動的ネットワークに対して適用するのは困難である．この問題に対し，新規エッジの隣接ノード集合を用いることで効率的に出現位置の推定ができることを示した．また後者について，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手法を提案した．人工ネットワークと実データに対して提案手法を適用し，エッジの影響力を分析・推定した．その結果，情報拡散の性質をもつネットワークにおいて，ネットワークを周縁部で強化するエッジが成長を誘発する傾向にあることを示した．

研究指導教員：佐藤 哲司
副研究指導教員：芳鐘 冬樹

目次

図目次

第 1 章

はじめに

1.1　動的ネットワーク

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混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 [1] や平均ノード間距離 [2]，次数分布のべき指数 [3]，有向ネットワークの指標としてモチーフ [4] 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 [5] 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「ネットワークを強化したか？ 拡張したか？」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，ネットワーク（コミュニティ）を強化する役割があるといえる．一方で，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，古参のネットワーク（コミュニティ）を拡張し，新参ファンを招き入れる役割があるといえる．これらの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．

1.2 研究の貢献

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ことである．

本研究の貢献は次の通りである．

- **ネットワーク上の位置の効率的な計算手法の提案**
  ネットワーク上での位置を求める指標としては，自身のノードと他のノードが平均的にどれくらい近いかを定量化した近接中心性 [5] が代表的である．しかし，この手法は自身と全てのノード間の最短距離を求める必要があり，動的ネットワークに対して計算するのは困難である．そこで，隣接ノード集合の規模を用いることで効率的にネットワーク上の位置を計算できることを示した．
- **効率的な成長誘発エッジ抽出手法の提案**
  エッジの出現時における「ネットワーク上の位置」と「ネットワークを拡張したか・強化したか」の 2 つの特徴量を用いて，ネットワークの成長をどの程度誘発するかを分析した．その結果，情報拡散の性質を持つネットワークにおいて，周縁部でネットワークを強化するエッジが，後続して出現するエッジを誘発する傾向にあることを明らかにした．

1.3 本論文の構成

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について説明する．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．

第 2 章

関連研究

本章では，動的ネットワークに関連する研究と本研究の位置づけを明らかにする．

2.1 情報カスケード

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 [6] 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．本節では，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 [7] 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 [8] 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を用いて，どのような特徴があるかを明らかにしている．

また，情報拡散モデルとしてよく用いられる独立カスケードモデル [9][10][11] と線形閾値モデル [12][13] 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．これを繰り返すことで，情報拡散をシュミ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ュミレートする．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 [14] 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ュミ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 [7] と吉川ら [14] 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としては，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目的変数とする点，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．

2.2 複雑ネットワークの生成モデル

本研究では評価実験に人工ネットワークを用いる．そこで，基本的な複雑ネットワークの生成モデルと実験に用いたモデルについて説明する．

まず，複雑ネットワークの代表的なモデルとして WS モデル [2] と BA モデル [15] が挙げられる．WS モデルは，スモールワールド性を満たすネットワークを生成するモデルとしてよく知られたネットワークである．スモールワールド性とは，いわゆる 6 次の隔たりと呼ばれるような，ネットワークのいずれのノードからでも少ないノードを介すだけでほぼ全てのノードに到達できるという性質である．また，BA モデルは動的なノードの追加と優先選択アルゴリズムによりスケールフリー性を実現したモデルである．スケールフリー性とは少数のノードが多くのノードからのリンクを集め，次数分布がべき乗になる性質である．ソーシャルネットワークなどで有名人が膨大なフォロワー数を持つのに対し，一般ユーザのそれはごくわずかである様子にあてはまる．その他，ソーシャルネットワークや友人ネットワークをモデル化した研究として CNN モデル [16] が提案されている．これはクラスター性と呼ばれる「友達の友達は友達になる」という性質を満たすモデルである．新たなユーザを追加する，あるいは「友達の友達は友達になる」エッジを追加する 2 つの処理を繰り返すことでネットワークを生成する．実際のソーシャルネットワークにおいては，人気のあるユーザーをフォローした後，そのユーザのフォロワーもフォローするという行動をモデル化したものだといえる．

本研究では，上述したモデルのうち CNN モデルを実験に用いる．

2.3 複雑ネットワークの動的分析

動的ネットワークを分析する研究として，時間経過に伴う構造変化の分析や，周期的な変化をみせるネットワークにおける異常値検出などが取り組まれている．

動的ネットワークの時間経過に伴う性質変化に着目した研究として，Leskovec[17] の研究がある．ここでは，時間経過とともに出次数が増大し直径が減少する性質が報告されている．また，Albert[18] はノードやエッジを削除する際のネットワークの頑健性を調査し，高次数ノードとエッジの削除がネットワークの平均クラスタリング係数と直径を大幅に変化させることを示した．Peel ら [19] は階層ランダムグラフモデル [20] を使用してコミュニティ階層のレベルを推定し，階層レベルが大幅に増加・減少するタイミングを検出している．Fushimi ら [21] は，エッジがネットワークに追加または削除されたときの各ノードの影響度を定量化することにより，構造変化の影響尺度を提案している．人工ネットワークと実際のネットワークを使用し，遠くのノード間のリンクの追加やコミュニティ間のリンクの削除など，意味のある変化を検出できることを示した．Koujaku[22] は，疎結合なノードが密に繋がったり，密に繋がっているノードが分離するなどの顕著な変化の検出を目的として，隣接行列の固有値分解を利用した動的ネットワークの異常値検出法を提案している．

第 3 章

問題設定：誘発スコア

本研究で取り組む問題について説明する．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．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よく用いられる線形閾値モデル [12] を応用し，**式 3.1** 及び**式 3.2** のようにエッジの**誘発スコア** $i_t(u,v)$ を定量化する．

図 3.1 を例に説明する．まず，**図 3.1a** 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 $e_2=(v_b,v_c)$ の出現を誘発する．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 3.1c** 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．

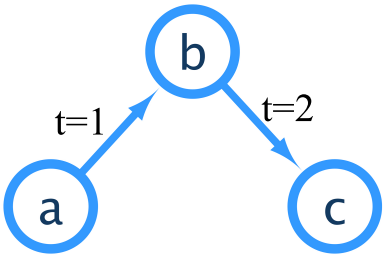
$$s_t((u,v)) = \frac{|\,\mathrm{OV}(v)|}{|\,\mathrm{IV}(v)|} \tag{3.1}$$

$$i_t((u,v)) = \begin{cases} 0 & (|\,\mathrm{OV}(v)|=0) \\ s_t(u,v) + \frac{1}{|\,\mathrm{IV}(v)|}(\Sigma_{x\in\mathrm{OV}(v)}\{s_t((v,x)) + i_t((v,x))\}) & (|\,\mathrm{OV}(v)|>0) \end{cases} \tag{3.2}$$

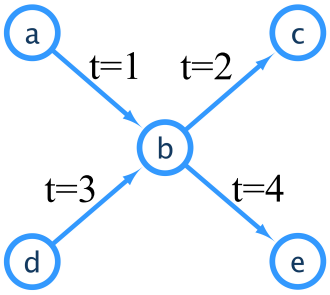
また，**図 3.1b** のようなケースにおいて上述した算出方法では，$e_1 = (v_a, v_b)$ の誘発スコアが 2 となってしまう．しかし，このままでは古いエッジほど誘発スコアが次々と累積され大きくなってしまう問題がある．これに対処するため，各エッジについてスコアを確定するタイミングを設ける．ノード v への入エッジの誘発スコアは，スコアが 0 を上回っている状態で，新たにノード v への入エッジが出現したタイミングで確定する．

図 3.1b 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 $i_1(v_a, v_b)$ は当然 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，e_1 のスコアが確定する．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．

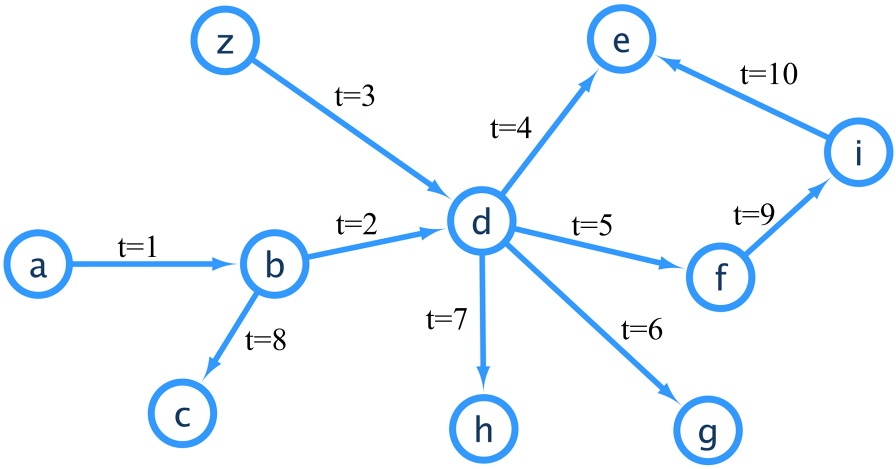
誘発スコアはネットワークの成長と共に累積していく値であるが，本研究ではこれをエッジ出現時の特徴から分析・推定することを目指す．



(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



$$i_9(v_f, v_i) = \frac{1}{1} + \frac{0}{1} = 1$$
$$i_5(v_d, v_f) = \frac{1}{1} + \frac{1}{1} = 2$$
$$i_2(v_b, v_d) = \frac{4}{2} + \frac{2+0+0+0}{2} = 3$$

(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 3.1: 誘発スコアの模式図

第 4 章

提案手法

1 章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，本節ではエッジの「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．

4.1 準備

本節では，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)[5] 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 $NV^{(t)}(u)$ と $NV^{(t)}(v)$ を用いて

式 **4.1** のように定義する．

$$n_t((u,v)) = \frac{|NV^{(t)}(u) \cup NV^{(t)}(v)|}{|V^{(t)}|} \tag{4.1}$$

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，本指標は次数中心性 [5] 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 **4.2**）．

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．

$$j_t((u,v)) = \frac{|NV^{(t)}(u) \cap NV^{(t)}(v)|}{|NV^{(t)}(u) \cup NV^{(t)}(v)|} \tag{4.2}$$

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 $j_t(u,v)$ 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．

第 5 章

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，Twitter のインタラクションデータセット，電子メールのやりとりのデータセットを用いている．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，ネットワーク名，ノード数，静的エッジ数，動的エッジ数について**表 5.1** に示す．実データ (Twitter のインタラクションデータセット，電子メールのやりとりデータセット) においては，同一の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．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 5.1: ネットワークのノード数及びエッジ数

データソース	ネットワーク名	ノード数	静的エッジ数	動的エッジ数
CNN ($p = 0.5, l = 1,000$)	CNN1K-NW	497	994	994
CNN ($p = 0.5, l = 10,000$)	CNN10K-NW	5,032	9,994	9,994
Higgs Twitter Dataset (Reply)	Reply-NW	1,233	1,622	1,971
Higgs Twitter Dataset (Mention)	Mention-NW	31,947	44,566	51,472
Higgs Twitter Dataset (Retweet)	Retweet-NW	45,804	62,817	74,380
email-EU-core-temporal Network	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[16] 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．CNN モデルの生成プロセスを Algorithm1 に示す．CNN は確率パラメータ p に基づき，4 行目から 7 行目の処理と 9 行目から 11 行目の処理のいずれかを繰り返す．前者の処理は新たなエッジとノードを追加する処理，後者の処理はポテンシャルリンクを実リンクへと変換する処理である．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，ループ回数毎にそれぞれ 10 パターン，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

Algorithm 1: CNN model algorithms

Data: CNN Network G, paramater q, loops l

Result: CNN Network G

for $k = 0;\ k < l;\ k++$ **do**
  set random value $r_k (0 \leq p \leq 1)$;
  if $r_k \leq p$ **then**
    ネットワーク G に新規ノード v_i を追加する
    既存ノードからランダムに一つノード v_j を選択する
    エッジ $e(j, i)$ を追加する
    v_j の隣接ノードと v_i を結ぶ無向エッジを**ポテンシャルリンク**に設定する
  else
    ポテンシャルリンクからランダムに 1 つ無向エッジ $e(u, v)$ を選ぶ
    選択したエッジの向きをランダムに決定する
    有向エッジ $e(u, v)$ か有向エッジ $e(v, u)$ を G に追加する

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．**表 5.1** には評価実験のうち個別に用いる 2 つを代表例として示した．

5.2 Higgs Twitter Dataset

Higgs Twitter Dataset[*1][23] 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 にリプライを送ると，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，インタラクション種類別に Reply-NW, Mention-NW, Retweet-NW の 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 5.1** に示す．Reply-NW はリプライツイートからなるネットワークである．ネットワークの規模は，他の実データから構築されたネットワークに比べると比較的小さい．同様に，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．Retweet-NW について，時刻 t にユーザ u のツイートをユーザ v がリツイートしたとき，情報の流れを表現するため，有向エッジは $e_t = (v, u)$ となる．なお，本研究では構築されたネットワークの最終時刻における最大連結成分を抽出し，実験対象とした．また，本章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．

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

5.3 email-Eu-core temporal network

email-Eu-core temporal network[*2][24] 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．

他のネットワークとは異なり，ノード数に対しエッジ数が非常に大きいデータセットとなっている．これは，研究機関に所属してメールをやりとりする人数が SNS ほど膨大にはならないことと，限られた人員の中で多くのメールがやり取りされるというデータセットの特徴に起因する．また，データの収集期間がおよそ 2 年半近くと長期間にわたることも動的エッジ数が膨大になっている理由の一つである．

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

第 6 章

評価実験

本章では 5 章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．評価に使用する隣接スコアと強化拡張スコア，誘発スコアの役割とそれら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 章で説明したように，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 の評価指標は規模毎に 10 パターンの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できているか評価するために，近接中心性との比較を行う．しかし，最もよく知られている Freeman による近接中心性 [5] 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 [25][26] 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こととする．

ノード集合を V，ノード u, v 間の最短経路長を $d(u, v)$ とすると，ノード u の調和近接中心性 $c^v(u)$ は**式 6.1** のように定義される．ノード u, v が非連結な場合，$d(u, v)$ は無限大であるので $\frac{1}{d(u,v)} = 0$ となる．これを用いて，エッジ $e(u, v)$ の調和平均近接中心性 $c^e(u, v)$ を**式 6.2** と定義する．また，本研究ではこれ以降，近接中心性に言及する際にはエッジの調和近接中心性を指しているものとする．

$$c^v(u) = \frac{\sum_{v \in E} \frac{1}{d(u,v)}}{|V| - 1} \tag{6.1}$$

$$c^e((u, v)) = \frac{c^v(u) + c^v(v)}{2} \tag{6.2}$$

6.1.2 近似の妥当性

5 章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，ピアソンの相関係数，$nDCG$，スピアマンの相関係数の 4 つを用いた．以下に，それぞれについて詳細を述べる．

6.1.2.1 近接中心性上位 (下位) エッジを隣接スコアで抽出できるか

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．これは，ランキングとしての妥当性を検証することがねらいである．また，近接中心性のスコアが上位であれば中心部に，下位であれば周縁部に位置することを意味するため，上位下位の両方を検証する．ここで，エッジ $e = (u, v)$ について近接中心性を返すスコア関数を f_c, 隣接スコアを返すスコア関数を f_n，エッジ集合 E からスコア関数 f 上位 $x\%$ のエッジを抽出した集合を $top_x(E, f)$ とすると，再現率 $Recall@top_x$ は**式 6.3** のように定義される．また，下位についても同様にスコア関数 f 下位 $x\%$ のエッジを抽出した集合を $worst_x(E, f)$ とすると，再現率 $Recall@worst_x$ は**式 6.4** と定義される．なお，このケースでは正解エッジ数と抽出エッジ数は常に等しいため，再現率は精度と等しい．

$$Recall@top_x = Precision@top_x = \frac{|top_x(E, f_n) \cap top_x(E, f_c)|}{|top_x(E, f_n)|} \tag{6.3}$$

$$Recall@worst_x = Precision@worst_x = \frac{|worst_x(E, f_n) \cap worst_x(E, f_c)|}{|worst_x(E, f_n)|} \tag{6.4}$$

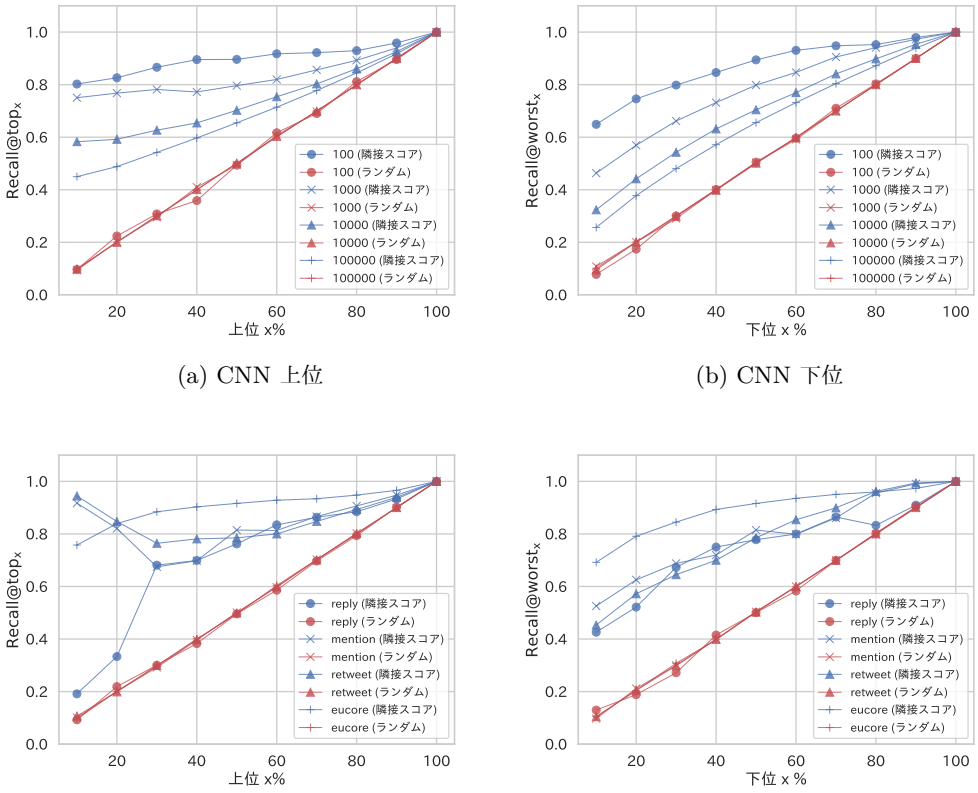
人工データと実データに対して提案手法を適用し，上述した再現率を求めた．対象は CNN4 種 (ループ回数 $l = 100, 1,000, 10,000, 100,000$)，Reply-NW, Mention-NW, Retweet-NW, Eucore-NW，合わせて 7 つのネットワークである．また，ベースラインとしてランダムにエッジを抽出した際の再現率も併せて算出した．**図 6.1** に正解エッジ割合に対する再現率の推移を示す．横軸が正解エッジの割合，縦軸が再現率，青いプロット線が提案手法に，赤い線がランダム抽出に対応する．

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 6.1c**, **図 6.1d** 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 ポイント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．

6.1.2.2 スコアの分布は妥当か

隣接スコアと近接中心性の分布を比較するために，これらの相関関係を調査する．

まず，隣接スコアと近接中心性の散布図を**図 6.2** に示す．対象データは CNN1K-NW, CNN10K-NW, Reply-NW, Mention-NW, Retweet-NW, Eucore-NW とする．横軸が近接中



(a) CNN 上位　(b) CNN 下位

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

図 6.1: 上位/下位 x% 抽出時の再現率

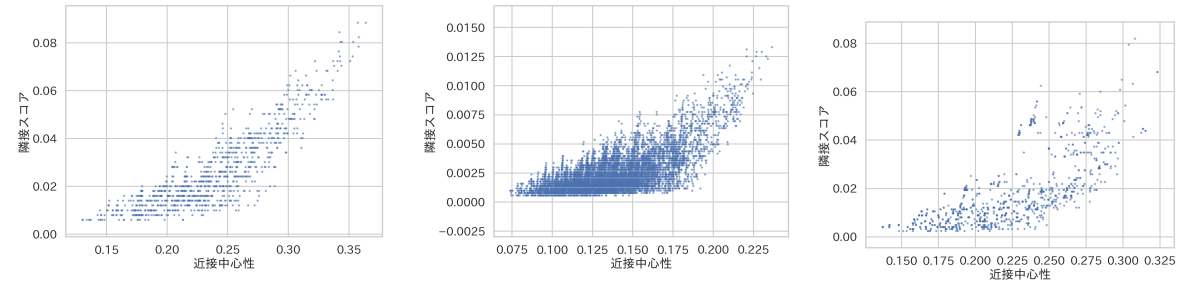
心性，縦軸が隣接スコア，各点がエッジに対応している．近接中心性が大きいほどネットワークの中心部に位置することを意味し，隣接スコアが大きいほどエッジ両端の隣接ノード集合の和集合のサイズが大きいことを意味する．いずれのネットワークにおいても，正の相関関係が観察できる．また，人工ネットワークである CNN と実データの間にも，分布の形状に大きな乖離は見られない．

次に，定量的な評価を行うために隣接スコアと近接中心性について，ピアソンの相関係数を測る．CNN 全 12 種と Reply-NW, Mention-NW, Retweet-NW，Eucore-NW に対して実験を行った．それぞれの相関係数の推移を**図 6.3** に示す．CNN についてはネットワークの種類が多いため，横軸にネットワークの規模を取り，折れ線グラフとした (**図 6.5a**)．縦軸が相関係数，青い線 (バー) がピアソンの相関係数に対応する．6.1.2.1 項と同様，ネットワークの規模が大きくなるほどスコアが下がる傾向にあるものの，いずれにおいても，「強い相関」から「やや相関あり」となった．このことから，分布という観点においても，提案手法は近接中心性と同様の性質を持っており，ネットワーク上の位置を定量化できているといえる．

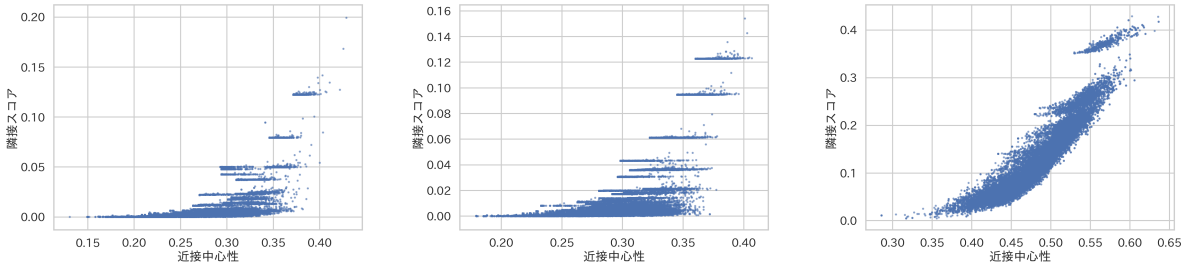
6.1.2.3　ランキングとして妥当であるか

上述した，6.1.2.1 項では，上位 (下位) のエッジを正しく推定できるかを確かめた．また，6.1.2.2 項では，隣接スコアと近接中心性が直線的な相関関係を持つか確かめた．

本節では，スコア全体のランキングとしての妥当性を評価するため，nDCG(normalized Discounted Cumulative Gain)[27] 及びスピアマンの相関係数を算出した．nDCG はランキン

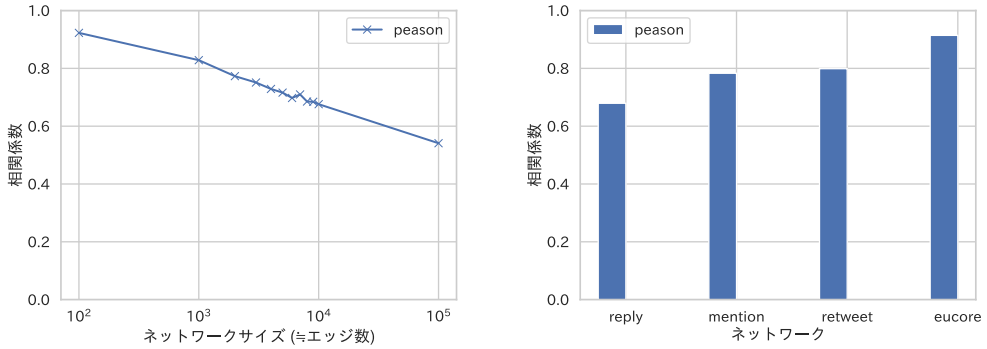


(a) CNN1K-NW (b) CNN10K-NW (c) Reply-NW



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

図 6.2: 隣接スコアと近接中心性の散布図

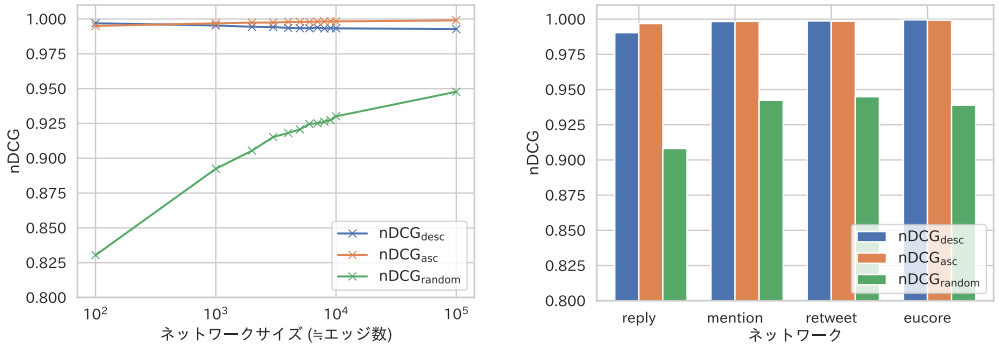


(a) CNN (b) Higgs と Eucore

図 6.3: 隣接スコアと近接中心性のピアソンの相関係数

グの評価指標であるが，実際のスコア (ここでは近接中心性) を考慮する指標である．予測手法に基づきランキングした正解データのスコアの和 (DCG) を正規化したものであり，大きいほどランキングが正しいことを示す．ただし，ランキングが低いほどスコアが小さくなるよう係数がかかる．すなわち，スコアが高いもの (本来ランク上位のもの) を低ランクにしてしまった時ほどペナルティが大きくなる．

ランキング i 番目のスコア (近接中心性) を r_i，ランキングするアイテム (エッジ) 数を N とすると，DCG は**式 6.5** のように定義される．また，**式 6.6** のように，予測データに基づく DCG を正解データの $DCG_{perfect}$ で割ることで，正規化された nDCG が得られる．ここでは，近接中心性をランキングした $DCG_{c(e)}$ が $DCG_{perfect}$ にあたる．nDCG はスコアを大きい順に並べる必要があるが，近接中心性はその大小いずれにも意味がある．そのため，昇順と



(a) CNN　　(b) Higgs と Eucore

図 6.4: $nDCG$ の比較

降順の両方で実験を行った．このとき，正解データとして隣接スコアを降順に並べた場合は $r = c(e)$ を，隣接スコアを昇順に並べた場合には $r = 1 - c(e)$ を用いた．

$$DCG = r_1 + \sum_{i=2}^{N} \frac{r_i}{log_2 i} \tag{6.5}$$

$$nDCG = \frac{DCG_{n(e)}}{DCG_{perfect}} = \frac{DCG_{n(e)}}{DCG_{c(e)}} \tag{6.6}$$

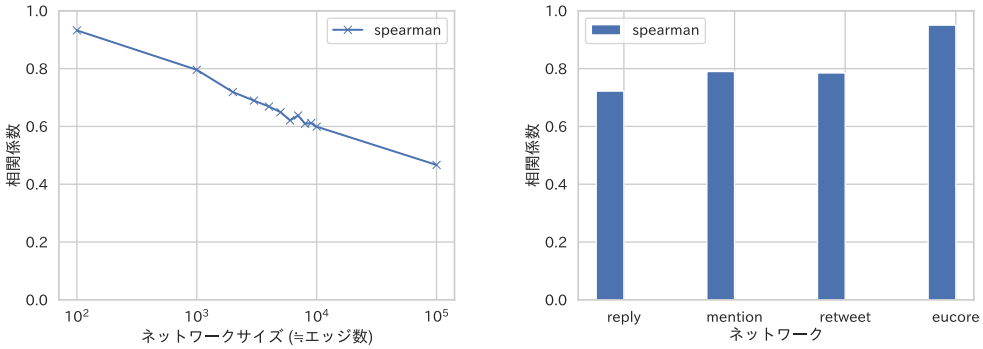
CNN 全 12 種と Reply-NW, Mention-NW, Retweet-NW，Eucore-NW の計 16 ネットワークの隣接スコアと近接中心性について nDCG を算出した結果を**図 6.4** に示す．併せて，ランダム抽出の場合の nDCG も算出している．**図 6.3** 同様，CNN については折れ線グラフで示してある．縦軸は $nDCG$，降順の $nDCG_{desc}$ が青い線 (バー)，昇順の $nDCG_{asc}$ がオレンジの線 (バー)，, 緑の線（バー）がランダム抽出の $nDCG_{random}$ に対応している．まず，ベースラインとしたランダム抽出は最も高いスコアの Eucore-NW でも 0.950 程度であり，Reply-NW は 0.900 を下回る結果となった．一方，すべてのネットワークについて昇順・降順いずれにおいても nDCG は 0.975 以上を示している．

また，**図 6.5** に，nDCG と同様に隣接スコアと近接中心性についてスピアマンの相関係数を算出した結果を示す．規模が大きくなるほどスコアが下がる傾向にあるものの，すべてのネットワークで相関係数 0.5 0.9 程度であり，中程度から強い相関があることを示している．

これらのことから，提案手法は近接中心性を近似できており，ネットワーク上の位置を定量化できる手法だといえる．

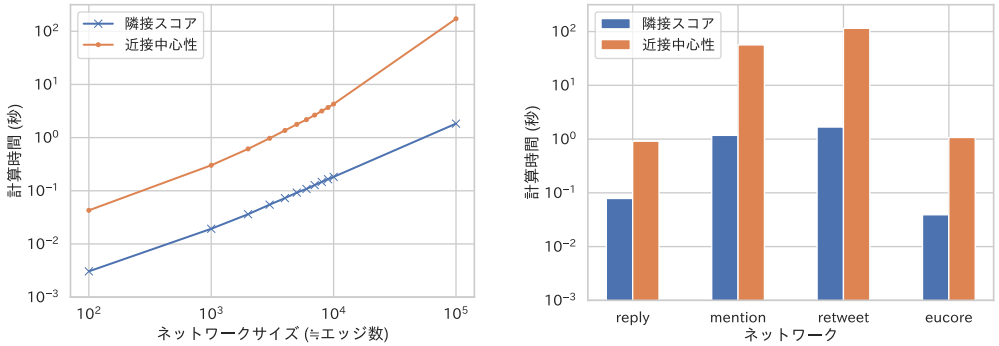
6.1.3　計算効率

5 章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，両指標の計算時間を**図 6.6** に示す．**図 6.3**, **図 6.4** 同様，CNN については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



(a) CNN　　(b) Higgs と Eucore

図 6.5: 隣接スコアと近接中心性のスピアマンの相関係数



(a) CNN　　(b) Higgs 及び Eucore

図 6.6: 隣接スコアと近接中心性の計算時間比較

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．

6.2　隣接スコアと強化拡張スコアから誘発スコアを推定できるか

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 は規模毎に 10 パターン存在するが，本節では**表 5.1** 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是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 5.1** 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．

分析の概要は次のとおりである．まず，各ネットワークについて概観するため，隣接スコア，強化拡張スコア，誘発スコアの 3 指標の分布を確認した．次に，提案手法の有効性を確認するため，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．最後に，一切のエッジを誘発しない誘発スコア $i(e) = 0$ のエッジに関する分析を行った．

6.2.1 各指標の分布

本節では，隣接スコア，強化拡張スコア，誘発スコアの 3 指標の分布を確認する．隣接スコアと強化拡張スコアはエッジ出現時に，誘発スコアは全てのエッジが出現した後 (最終時刻) に算出を行った．それぞれについて以下に示す．

6.2.1.1 隣接スコア

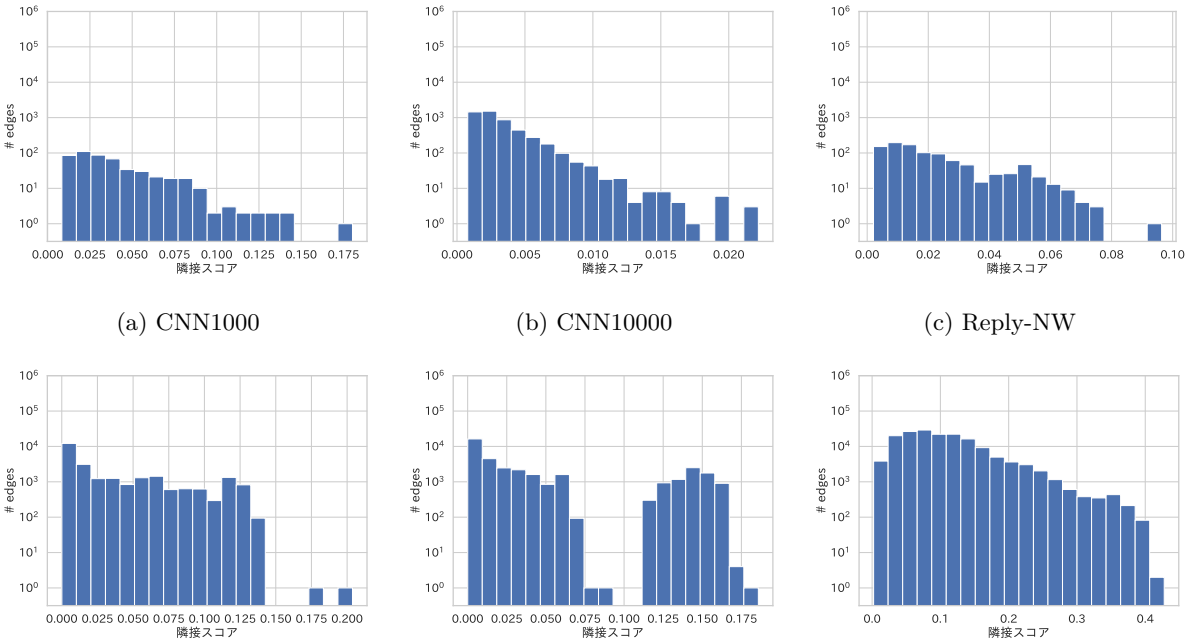
各ネットワークにおける隣接スコアのヒストグラムを**図 6.7** に示す．横軸が隣接スコア，縦軸がエッジ数に対応している．

人工データである CNN（**図 6.7a** と**図 6.7b**）は，べき乗則に従い分布している．これに類似する実データとして，**図 6.7c** の Reply-NW や**図 6.7f** の Eucore-NW が挙げられる．これらのネットワークでは，隣接スコアの値が小さい（ネットワーク周縁部に出現する）エッジが多くを占めるといえる．一方，Mention-NW と Retwet-NW の振る舞いはいずれも大きく異なっている．**図 6.7d** の Mention-NW は一様分布に近い形を示しており，中心から周縁に渡って偏りなくエッジの出現が観測されること意味する．また，**図 6.7e** の Retweet-NW においては，エッジの出現位置が二分されている．隣接スコアが 0 から 0.075 付近までは，ややべき乗則に従うような分布を示し，そこから間を開けて一つの大きな山を形作っている．この右側の山は，極端に多くのリツイートを集めたユーザによるものだと考えられる．Retweet-NW は，ツイートをしたユーザとリツイートをしたユーザの間にエッジが付与されるネットワークである．そのため，多くのリツイートを集めたユーザとリツイートしたユーザの間で 1 対多の関係が構築される．この時，これらの関係を構築するエッジの隣接ノード集合の和の規模は比較的近いものになる．そのため，類似するスコアのエッジが多く発生していると考えられる．

6.2.1.2 強化拡張スコア

各ネットワークにおける強化拡張スコアの分布を**図 6.8** に示す．横軸が強化拡張スコア，縦軸がエッジ数に対応している．強化拡張スコアが大きいほどリンク相手との類似度が高い強化型のエッジ，小さいほどリンク相手との関係が少ない拡張型のエッジであることを示す．各ネットワークに共通する傾向として，強化拡張スコアが小さいエッジが多くを占めることが挙げられる．特に，**図 6.8d** の Mention-NW と**図 6.8e** の Retweet-NW の分布がべき乗則に従っており，その傾向がより強く表れているといえる．このことから，Mention-NW と Retweet-NW はネットワークの拡張を中心に成長しているといえる．

また，**図 6.8f** の Eucore-NW では，0.0~0.4 あたりにかけて偏りなく分布していることが分かる．Eucore-NW はノード数に対してエッジ数が非常に多いため，時間が経つにつれてネットワークが密になる．その結果，共隣接ノード集合が大きくなり，強化拡張スコアの高いエッジが出現していると考えられる．



(a) CNN1000　(b) CNN10000　(c) Reply-NW

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

図 6.7: 隣接スコアの分布

6.2.1.3　誘発スコア

各ネットワークにおける誘発スコアの分布を**図 6.9** に示す．全ネットワークに共通する傾向として，誘発スコアの分布がべき乗則に従っている点が挙げられる．また，**図 6.9d** の Mention-NW，**図 6.9c** の Retweet-NW は，誘発スコアの小さいエッジが非常に多い．また，**図 6.9f** の Eucore-NW は．他のネットワークに比べ文字通り桁違いの誘発スコアを示す．これは，そもそもエッジの絶対量が多いためだと考えられる．

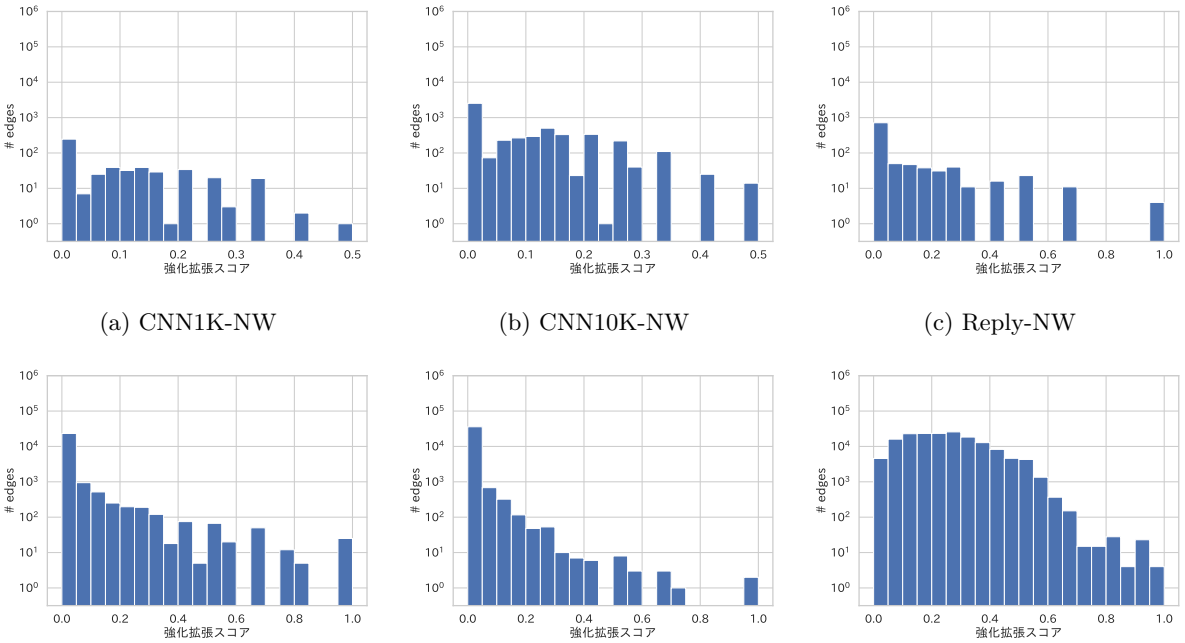
6.2.2　隣接スコア，強化拡張スコア，誘発スコアの関係

本節では，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 6.10** 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．この処理については後述する 6.2.3 項で分析結果に大きな影響を与えないことを示す．

図 6.10d の Mention-NW と**図 6.10e** 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あるといえる．

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



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

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

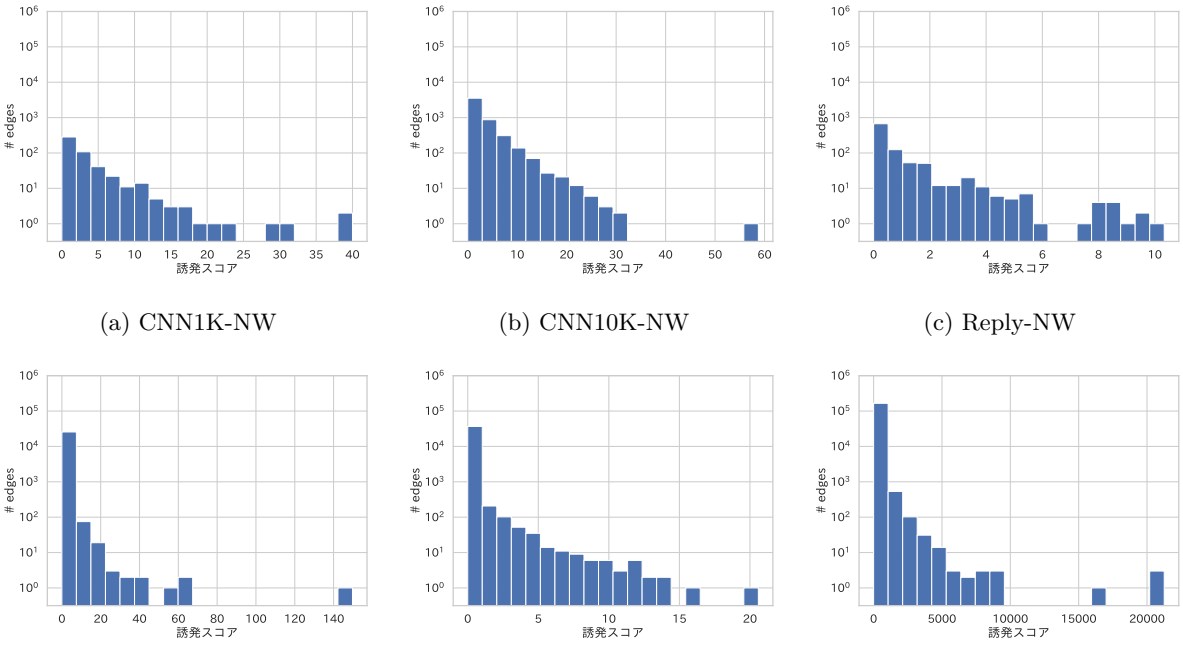
図 6.8: 強化拡張スコアの分布

表 6.1: 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

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 6.1** 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 6.3 節で考察する．



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

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

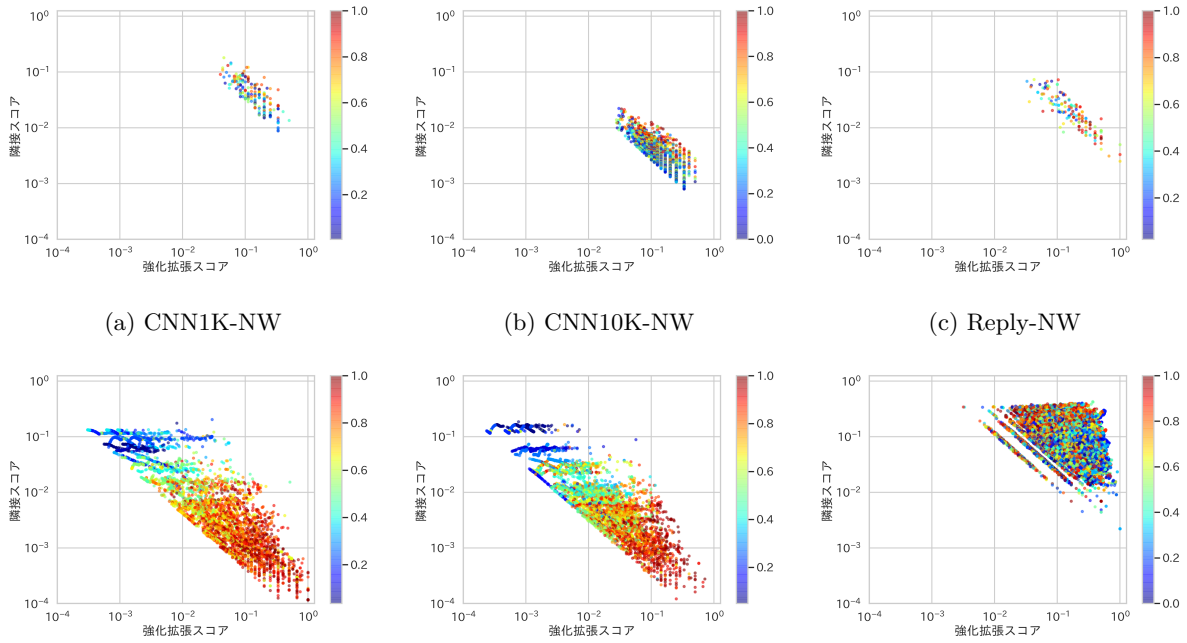
図 6.9: 誘発スコアの分布

6.2.3 誘発スコア $i(e) = 0$ のエッジを事前に検出できるか

6.2.2 項で隣接スコア，強化拡張スコア，誘発スコアの関係を分析する際に誘発スコアが 0 のエッジを分析対象から除外した．本来，誘発スコアは最終時刻にならないと判明しない未知の指標であるため，予測モデルの構築に伴う前処理としてはやや不適切である．そこで，Mention-NW と Retweet-NW の誘発スコア $i(e) = 0$ のエッジを含む散布図を図 **6.11** に示す．濃紺で示される誘発スコア $i(e) = 0$ のエッジは，他のエッジと大きく異なる振る舞いを見せており，これがモデルの性能を下げることは明らかである．これらのエッジをスコア確定以前に何らかの方法で除去することのメリットは大きい．そこで，「エッジ出現後しばらく誘発スコアが 0 のエッジは，最終時刻でも誘発スコアは 0」という単純なモデルについて検討する．そもそも誘発スコアは他者に与える影響力を定量化したスコアである．一般にソーシャルネットワーク等においては，発言の影響力は発信直後が最も高く時間経過とともにその影響は減衰していく．そのため，出現直後の誘発スコアを見ることで，将来他者を誘発するかしないか明らかにできると考えた．

実験の手順は次のとおりである．エッジ e の時刻 t における誘発スコアを $i_t(e)$ と表し，最終時刻の誘発スコアを $i_T(e)$ とする．誘発スコアの初期値は 0 であるので，誘発スコア $i_t(e)$ を求めることでどの時刻で誘発スコアが発生したのか（あるいは発生していないのか）がわかる．時刻 t において $i_t(e) = 0$ の場合，エッジ e の最終時刻 T における誘発スコアは $i_T(e) = 0$ と予測する 2 クラス分類モデルを構築する．逆に $i_t(e) > 0$ のときには，$i_T(e) > 0$ と予測する．つまり，エッジ出現後に一定時間経過した際のクラスが，最終時刻のクラスとなる単純なモデルである．複数の時刻で予測を行い，モデルの精度を確かめる．

また，データセットにおいて時刻 t はエッジリストのインデックスとして与えられている．しかし，各エッジの出現時点から単純に t をインクリメントしていくと，ネットワークの規模



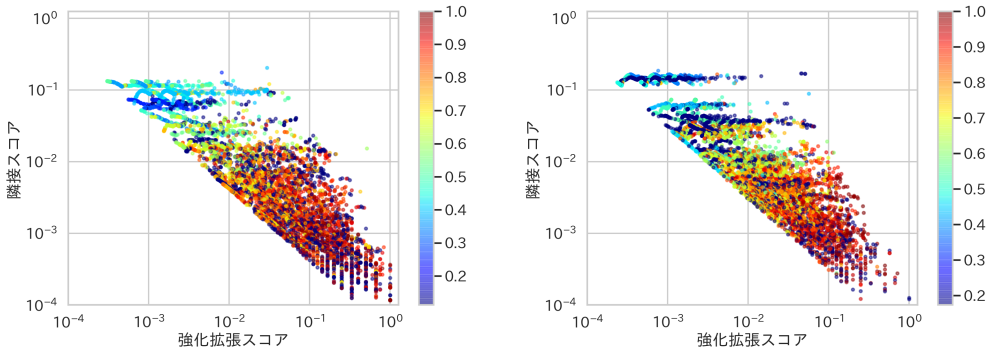
(a) CNN1K-NW (b) CNN10K-NW (c) Reply-NW

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

図 6.10: 隣接スコア，強化拡張スコア，誘発スコアの関係 ($(i(e) > 0)$)

の影響を大きく受ける．例えば，時刻 $t = 10$ に 10 エッジ目として追加されたエッジについて次の時刻 $t = 11$ の誘発スコアをモデルへの入力とする場合を考える．この時，時刻 $t = 1000$ の 1000 エッジ目についても同様に，次の時刻 $t = 1001$ の誘発スコアを予測値とすることは望ましくない．なぜなら，大規模なネットワークほどネットワークの各地で同時多発的に成長が起こっているためである．そこで，各エッジの出現時点におけるグラフの規模に基づきモデルへの入力時刻を決定することにした．時間経過の係数を定め，エッジ出現時点でのネットワーク中のエッジ数に時間経過係数を乗じる．時間経過の係数は，1.01 から 1.30 まで 0.01 刻みの 30 種とした．

$t = 100$ に出現したエッジの場合は時刻 $t = 101, 102, ..., 130$ の各時刻において予測を行う．また $t = 1,000$ に出現したエッジなら時刻 $t = 1,010, 1,020, ..., 1,300$ の誘発スコアを入力とする．モデルが $i_T(e) = 0$ と予測したもののうち実際に $i_T(e) = 0$ であったものの割合，すなわち Precison を**図 6.12** に示す．横軸が時間経過係数，縦軸が Precison，各プロット線がネットワークに対応する．まず，ネットワーク全体として時間経過とともに Precision が上昇する．これは上述した「エッジ出現後しばらく誘発スコアが 0 のエッジは，最終時刻でも誘発スコアは 0」という仮説を支持するものである．特に重回帰モデルの性能が良かった Mention-NW に着目すると，出現直後の 1.01 倍時点で 4 割弱，1.30 倍時点で 6 割弱 $i_T(e) = 0$ のエッジを抽出できている．また，Reply-NW については出現直後で Precision0.4 強，1.30 倍時点で 0.6 強を示す．Eucore-NW に関しても出現直後こそ低いものの，時間経過に従って急激に Precison を伸ばしている．単純なルールベースモデルでも一定の Precision が出ていることから，最終時刻において誘発スコア $i_T(e) = 0$ のエッジを事前に除外することは十分可能であるといえる．



(a) Mention-NW　　(b) Retweet-NW

図 6.11: 隣接スコア，強化拡張スコア，誘発スコアの関係 ($i(e)\geq 0$)

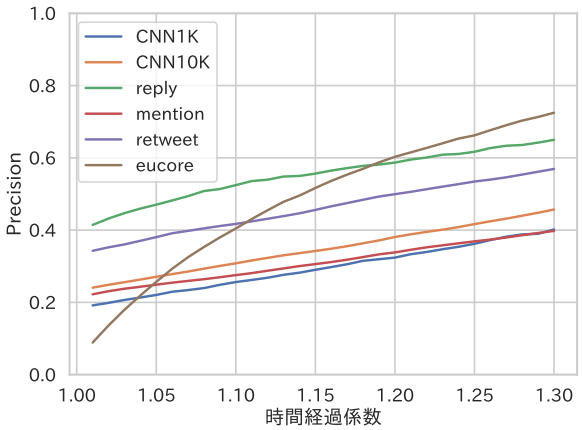


図 6.12: ルールベースモデルの Precision

6.3　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 6.10** で顕著に傾向が表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あ他者に与えられる可能性は高まる．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．**図 6.8** に示すように Mention-NW と Retweet-NW は拡張型のネットワークである．加えて，メンションやリツイートは他者に情報を伝える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

なコミュニティが出現・成長し，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）．これらの要因から，Reply-NW や Eucore-NW については，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，情報拡散のシュミ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．

また，今回の実験では重回帰分析を行う際に，誘発スコアが 0 のエッジを除外する処理を行っている．6.2.3 節でこの処理の妥当性について評価を行ったところ，非常に単純なルールベースモデルで 4~6 割近くの誘発スコアが 0 のエッジを検出することができた．このことから，ネットワーク構造などのより多様な特徴量を用いることで，高精度なモデルの構築が可能だと考えられる．

第 7 章

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，情報拡散をシュミ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

指導教員の佐藤哲司教授をはじめ，芳鐘冬樹教授，東京工科大学の伏見卓恭助教には多岐にわたり熱心なご指導をいただきました．ここに記して，感謝の意を申し上げます．また，院生室で学生生活の多くを共に過ごしたコミュニケーション理解研究室の小邦くんや小久保さんにも感謝いたします．

本研究に関連する発表論文

国内論文誌（査読付き）

[1] 稲福 和史, 伏見 卓恭, 佐藤 哲司, “トライアド推移に基づく購買行動の成長分析,” Japanese, **情報処理学会論文誌**, vol. 60, no. 4, pp. 1141–1150, Apr. 2019.

国際会議論文（査読付き）

[1] K. Inafuku, T. Fushimi, and T. Satoh, “Structural transition analysis of dynamic network based on roles of adding edges,” in *Proceedings of the 21th International Conference on Information Integration and Web-based Applications and Services*, ser. iiWAS ’19, Munich, Germany: ACM, 2019.

参考文献

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

[2] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

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

[4] 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, Oct. 2002.

[5] L. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[6] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.

[7] 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*, ACM, 2014, pp. 925–936.

[8] 川本貴史, 豊田正史, 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," Japanese, 第 *8* 回 *Web* とデータベースに関するフォーラム論文集, vol. 2015, pp. 48–55, 2015.

[9] 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*, vol. 12, no. 3, pp. 211–223, 2001.

[10] 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*, ACM, 2003, pp. 137–146.

[11] 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)*, vol. 3, no. 2, pp. 1–23, 2009.

[12] D. J. Watts, "A simple model of global cascades on random networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 5766–5771, 2002.

[13] D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," *Journal of consumer research*, vol. 34, no. 4, pp. 441–458, 2007.

[14] 吉川 友也, 斉藤 和巳, 元田浩, 大原剛三, 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,” Japanese, **電子情報通信学会論文誌** *D*, vol. 94, no. 11, pp. 1899–1908, 2011.

[15] A. L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[16] A. Vázquez, “Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations,” *Physical Review E*, vol. 67, no. 5, pp. 056104+, May 2003.

[17] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2–es, Mar. 2007.

[18] R. Albert, H. Jeong, and A. L. Barabási, “Error and attack tolerance of complex networks,” *Nature*, vol. 406, pp. 378–382, 2000.

[19] L. Peel and A. Clauset, “Detecting change points in the large-scale structure of evolving networks,” *CoRR*, vol. abs/1403.0989, pp. 2914–2920, 2014.

[20] A. Clauset, C. Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” *Nature*, vol. 453, pp. 98–101, 2008.

[21] T. Fushimi, T. Satoh, K. Saito, and K. Kazama, “Comparison of influence measures on structural changes focused on node functions,” in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, ser. iiWAS ’15, Brussels, Belgium: ACM, 2015, 16:1–16:10.

[22] S. Koujaku, M. Kudo, I. Takigawa, and H. Imai, “Structual Change Point Detection for Evolutional Networks,” in *Proceedings of the World Congress on Engineering*, vol. 1, 2013, pp. 324–329.

[23] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi, “The anatomy of a scientific rumor,” *Scientific reports*, vol. 3, p. 2980, 2013. [Online]. Available: `https://www.nature.com/articles/srep02980`.

[24] 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*, ser. WSDM ’17, Cambridge, United Kingdom: ACM, 2017, pp. 601–610.

[25] Y. Rochat, “Closeness centrality extended to unconnected graphs: The harmonic centrality index,” Tech. Rep., 2009. [Online]. Available: `https://infoscience.epfl.ch/record/200525`.

[26] P. Boldi and S. Vigna, “Axioms for centrality,” *Internet Mathematics*, vol. 10, no. 3-4, pp. 222–262, 2014.

[27] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.