

はじめに

概要

概要

まとめ

【論文紹介】

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方法

武藤 克弥 (Katsuya Mutoh)
u255018@st.pu-toyama.ac.jp

富山県立大学 大学院 電子・情報工学専攻 情報基盤工学部門

June 14, 2023

背景

- インターネット, SNS, 道路交通網などの現実のネットワークは日々ノード (頂点) とエッジ (辺) が変化する動的な構造である.
- 動的ネットワークで影響度の高くなるノードやエッジを特定できれば SNS マーケティングやコミュニティの成長予測につながる.

目的

- ネットワーク構造を測定する指標を新たに提案し, ネットワークの成長を誘発するエッジを抽出する.

情報カスケード

- 情報が1人から不特定多数に伝搬する現象が連鎖的に生じている現象.
- 正確にモデル化することで, 現実ネットワークを再現. 的確なマーケティングや成長予測に活かせる.

既存研究

- Facebook の写真共有機能で情報カスケードの規模を定量化し、情報拡散の連鎖が成長するかを予測
- 情報送信者を軸にしたカスケードモデルを構築し、SNS 上でどれくらいの人に情報が伝わったかを定量化

※ノードからノードへの情報拡散に注視して分析

本研究の新規性

- ネットワーク自体の成長に着目
- 情報発信を誘発する影響力のあるエッジを推定

SNS 上での情報カスケード (誘発スコアの概要)

- ノード (ユーザ) からノードに対してコメントやメールすることで、ユーザ間につながり (エッジ) が形成される
- 情報カスケードの中であるエッジが将来的にどれくらい他のエッジを誘発するのかを測定する

誘発スコア

情報拡散現象のモデル化に用いられる線形閾値モデルを応用

→提案手法の指標と組み合わせてネットワークを分析する

- $i_t(u, v)$: ノード u から v への誘発スコア
- $|OV(v)|$: ノード v に入るエッジ数, ノード v から出ていくエッジ数
- $s_t(u, v) = \frac{|OV(v)|}{|IV(v)|}$: エッジの出入り比率

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

ステップ $t = 2$ におけるエッジ $b \rightarrow d$ の誘発スコア $i_2(b, d)$

$$i_2(b, d) = s_2(b, d) + \frac{\sum_{x \in OV(d)} i_t(d, x)}{|IV(d)|} \quad (2)$$

$$= \frac{4}{2} + \frac{2 + 0 + 0 + 0}{2} = 3$$

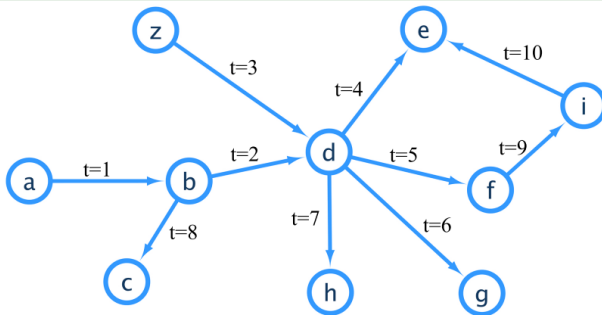


図 1: エッジの誘発

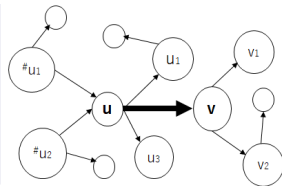
隣接スコア (中心性指標)

- つながりの多いノードほど影響力が強い＝中心性が高い
- 様々な中心性指標のうち隣接スコアは (巨大 net でも) 計算量が少ない
→ 影響力のあるノードからエッジが出現しやすい＝エッジの出現位置を測る指標
- $e = (u, v)$: エッジ (ユーザ u から v への情報伝達)
- $E^{(t)} = \{e_1, e_2, \dots, e_t\}$: 時刻 t までのノード間の情報伝達の集合
- $V^{(t)}$: 動的エッジ集合 (t までに追加されたノードの集まり)
- $NV^t(v) = \{(u, v) \in E^t \vee (u, v) \in E^t\}$:
隣接ノード集合 (ノード u に隣接するノードの集まり)

$n_t(u, v)$: エッジ $e_t = (u, v)$ の隣接スコア

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

$$= \frac{5 + 3}{12} \doteq 0.67$$



強化・拡張スコア

エッジの役割を測る指標

- エッジ両端のノードが持つ隣接ノード集合の Jaccard 係数と同義
- $j(u, v)$: エッジ $e = (u, v)$ の強化拡張スコア

$$j(u, v) = \frac{|NV^{(t)}(u) \cap NV^{(t)}(v)|}{|NV^{(t)}(u) \cup NV^{(t)}(v)|} \quad (3)$$

1 に近い : 隣接ノード重複が多い = ネットワークを強化するエッジ

0 に近い : 重複少ない = ネットワークを新たに拡張するエッジ

実験の概要

6 種類のネットワークデータセット (ND) を使用

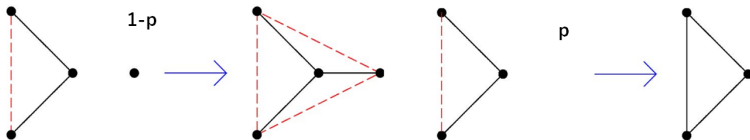
- ① 隣接スコアが，従来の中心性指標に比べて効率的な指標か評価
- ② 提案した隣接スコア・強化拡張スコアに対して誘発スコアに関係性が見いだせるか ND を基に分析

(1) Connecting Nearest Neighbor (CNN) モデル

- 「友達の友達は友達になる」の仮定で作られたネットワーク成長モデル
- 現実ネットワークの多くにみられる確率分布の挙動になる, 現実に近いモデル

時刻 t で以下のどちらかを選択してネットワークを成長

- 確率 $1 - p$ で新ノード追加. 既存ノードの 1 つとエッジを結ぶ.
→ 結んだノードと繋がっているノード全てにポテンシャルエッジを作る
- 確率 p でポテンシャルエッジから 1 つ選んで本物エッジにする



※ $p = 0.5$, ステップ数 $l = 1000$ と
 $l = 10000$ のデータセット (CNN1K-NW, CNN10K-NW) を使用

(2)Higgs Twitter Dataset

2012/7/1～2012/7/7 ヒッグス粒子が発見されたときのツイートアクション
(リプライ, メンション, リツイート) を収集したもの

- **lhc, cern, boson, higgs** のキーワードを含んだツイート
- 時刻 t でユーザ v のツイートに, ユーザ u がリプライ・メンション・リツイートした際のエッジ $e_t = (u, v)$ が記録
- Reply-NW, Mention-NW, Retweet-NW のデータセットを作成

(3)email-Eu-core temporal network

ヨーロッパ各研究機関のメールのやり取りを収集したもの

- 時刻 t でユーザ u がユーザ v にメール送信したエッジ $e_t = (u, v)$ が記録
- データセット名 Eucore-NW

データセットの構成

- 同じノード間にエッジが付与されるのを許可=動のエッジ数

表 1: データセットの構成

ネットワーク名	ノード数	静的エッジ数	動のエッジ数
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

隣接スコアの妥当性評価

- 高速な隣接スコアで従来の中心性指標 (調和近接中心性 (HC)) を再現できるか？
- エッジ $e_t = (u, v)$ の調和近接中心性： $hc(u, v)$

$$hc(u, v) = \frac{1}{2} \left\{ \sum_{d(u, x) < \infty, y \neq x} \frac{1}{d(\mathbf{u}, x)} + \sum_{d(v, y) < \infty, y \neq x} \frac{1}{d(\mathbf{v}, y)} \right\} \quad (4)$$

※ $d(u, x)$ ：ノード u と隣接ノード間の距離

- HC スコア上位 (+下位) $x\%$ のエッジを正解エッジとする
- 隣接スコア $x\%$ のエッジをどれだけ抽出できるか、再現率を計算
→ランダムでエッジ抽出した場合と比較=上位 $x\%$ 正解だと $x\%$ 必ず正解する

評価実験1 結果

14/19

はじめに
概要
概要
まとめ

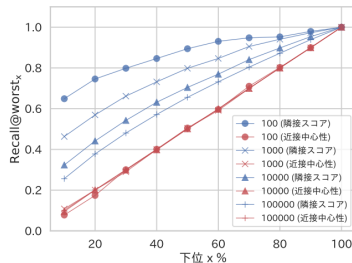
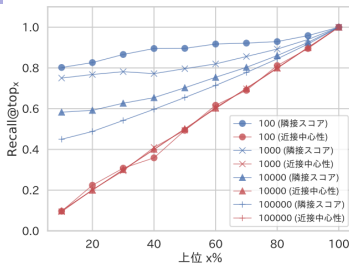


図 2: CNN 上位 $x\%$ (左) と下位 $x\%$ (右) の再現率

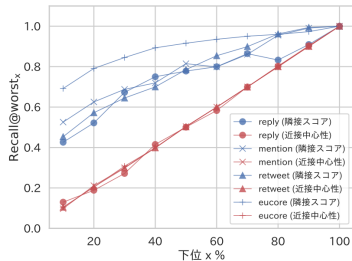
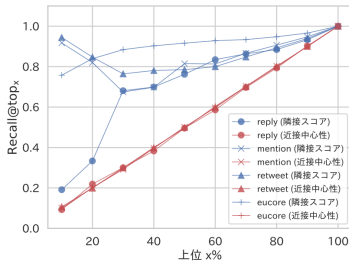


図 3: Higgs および Eucore 上位 $x\%$ (左) と下位 $x\%$ (右) の再現率

提案指標を用いたデータセットの分析

隣接スコアと強化拡張スコアによって誘発スコアが機能するか

- 表 1, 6 つのデータセットで分析
- データ不均衡の対策：
前半 25%, 後半 25%を除いた中央 50%だけ使用
(例. 前半=ノード少ない, 後半=誘発スコア 0 増加)
- 誘発スコアは 0-1 で正規化 (隣接・強化拡張スコアに合わせる)

横軸=強化拡張スコア. 縦軸=隣接スコア. 色 (青→赤)=誘発スコア.

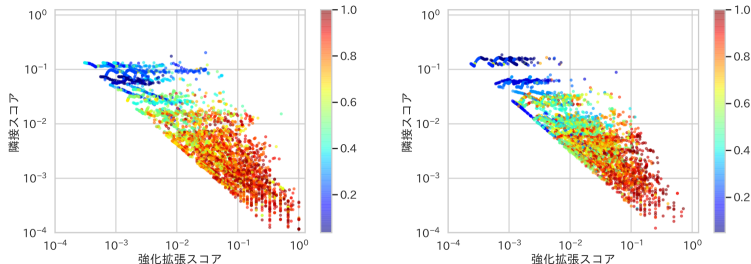


図 4: Mention-NW, Retweet-NW における 3 指標の関係

- 隣接スコア 低→強化拡張 高→誘発スコア 高
- 親密な相手とのコミュニティでネットワーク強化, エッジ誘発
→中心から離れた場所で新たな小規模コミュニティの出現
or 小規模コミュニティで情報拡散 (メンション & RT) が連鎖

- Eucore-NW(電子メール) や Twi リプライは情報拡散しにくい
=誘発スコアで成長を表現難
- 人工データの CNN-NW は確率パラメータ調整が必要

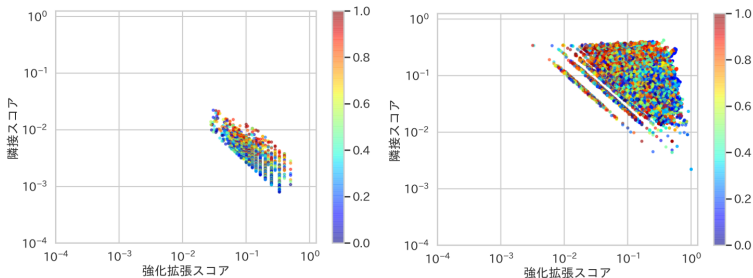


図 5: CNN10-NW, Eucore-NW における 3 指標の関係

重回帰分析

- 定量的な関係性を見る
- 隣接スコア, 強化拡張スコア=説明変数. 誘発スコア=目的変数

表 2: 重回帰分析結果

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

まとめ

- ネットワークの出現位置と役割を効率的に定量化し、ネットワークの成長を促すエッジを抽出する手法を提案
- Twitter データセットで小規模コミュニティを出現・強化させる傾向が示された

今後の課題

- より情報拡散に準拠した人工ネットワークを用いて検証
- 情報の内容に着目した評価指標 (特徴量) の追加

自分の研究に対して考えていること

- ネットワーク成長モデルをメカニズムデザインとして利用
- トピックの変遷=成長モデル として流行するトピックを予測する