L-005

リンク予測を用いたソーシャルネットワークデータの k-匿名化手法の検討
A k-Anonymization Method for Social Network data with Link Prediction

菅井 理紗[1)] 清 雄一[1)] 田原 康之[1)] 大須賀 昭彦[1)]
Risa Sugai Sei Yuichi Yasuyuki Tahara Akihiko Ohsuga

1 はじめに

近年，ソーシャルネットワークサービスの浸透により，ソーシャルネットワークデータの利活用への需要が高まっている. ソーシャルネットワークサービスを提供している組織は，ユーザの属性や行動に関する情報を含んだデータを蓄積しており，データの一部を第三者に提供する．データの提供を受けた第三者は，ユーザの所属するコミュニティや行動傾向について分析をすることによって，新サービスの開発や商品もしくはサービスの収益向上に役立てることができる．

しかし，ソーシャルネットワークデータにはプロフィールといった属性情報の他に，フォロー関係のような構造情報が含まれている．そのため，ソーシャルネットワークデータのプライバシを保護するには構造情報を考慮する必要がある．また，実際のソーシャルネットワークデータには欠損が含まれる可能性がある．実際，Facebook や Twitter といったソーシャルネットワークサービスではプロフィールやフォロー/フォロワーといった情報を非公開に設定することができる．

一般的に欠損値を含むデータセットでは欠損を取り除いた上で匿名化処理および分析を行うことが多い．しかし，欠損は匿名化後の情報損失量の増加に影響するため，実データへの匿名加工の適用を想定する場合は欠損の有無も考慮する必要がある．

本研究では，欠損値を補完することによる匿名加工データの有用性向上を目的とし，リンク予測を用いたソーシャルネットワークデータの k–匿名化手法の検討を行う. また，有用性評価として元データの特性を残しているかの確認を行うため，元データと一致するリンクの数を用いて評価を行う.

2 関連研究

2.1 リンク予測

ソーシャルネットワークデータはユーザをノード，フォローといったリンク情報をエッジとしてグラフデータと表現することができる．その際，無向ネットワークはノードの集合 **V**，エッジの集合 **E** で構成され，**G**=(**V**, **E**) と表現することができる．ソーシャルネットワーク分析において，グラフ理論を用いてリンクの存在可能性について予測するリンク予測手法が研究されている．

Common Neighbor and Centrality based Parameterized Algorithm(CCPA) は，2020 年に Ahmad ら [2] が提案したリンク予測手法である．CCPA スコアは共通の隣接と中心性に基づいて計算され, 次のように定義される．

$$CCPAscore = \alpha \cdot (|\Gamma(u) \cap \Gamma(v)|) + (1-\alpha) \cdot \frac{N}{d_{uv}} \quad (1)$$

$\Gamma(u)$，$\Gamma(v)$ は各ノードの近接ノードの集合，N はノードの総数，d_{uv} はノード間の最短距離である．CCPA スコアは大きい値となるほど，ノード u, v 間のリンクの存在可能性は高いとみなすことができ，しきい値以上の場合にリンクとして予測される．

2.2 k- 匿名化

プライバシ保護手法の 1 つに k-匿名化 [1] がある．k-匿名化とは，同一の属性を持つデータを少なくとも k 個以上存在させるようにデータ加工することによって，再識別のリスクを 1/k に低下させる手法である．組み合わせることで特定のデータを識別できる属性のことを準識別子と呼び，具体的には生年月日や性別，職業などが当てはまる．

また，グラフデータを対象とするプライバシ保護手法に k-degree[3] がある．k-degree は，ノードの次数を準識別子と設定し，同じ次数を持つノードが少なくとも k 個存在することを保証する匿名化指標であり，Liu らは k-degree を用いてエッジの追加または削除を行う匿名化手法を提案している．

たとえば，図 1 のような状況において，攻撃者は特定のユーザを指すノードの次数を背景知識として持ち，次数 2 のノードを特定したいと考えているとする．

(a) ソーシャルネットワーク　**(b)** 3-degree グラフ

図 1: k-degree の例

(a) における各ユーザの次数は {'Ada':2, 'Bob':1, 'Cathy':1} となる．このとき，次数 2 のノードは 1 つしかないため，(a) のグラフから Ada を一意に特定することができる．しかし，Bob と Cathy のリンクを新たに追加することで，(b) のグラフにおける各ユーザの次数は {'Ada':2, 'Bob':2, 'Cathy':2} となり，特定リスクは 1/3 に低下する．このとき，(b) のグラフは 3-degree を満たすと言える．

3 提案手法

3.1 攻撃者の背景知識の設定

本研究では，実データに欠損が含まれる場合を想定し，攻撃者が次の背景知識を持つと設定する．

- 特定したいユーザの次数に関する知識
- 元のデータセットがリンク予測されていること

攻撃者がリンク予測されていることを知っている場合，特定したいノード v の次数は実際の次数からリンク予測後のグラフ G' の最大次数までを範囲を持つ．

$$deg(v) \leqq deg'(v) \leqq \Delta(G') \quad (2)$$

1) 電気通信大学 大学院情報理工学研究科
The University of Electro-Communications Graduate School of Informatics and Engineering

攻撃者は上記の背景知識を用いてユーザの特定を行うものとする．

3.2 k-maximum_degree の定義

本研究では，リンク予測で欠損リンクを補完した後に匿名加工を行う手法について検討する．そこで，リンク予測後のデータを匿名化する状況において，匿名化指標 k-maximum_degree を提案する．

定義 グラフ G の次数列の最大値が少なくとも k 回出現するとき，グラフ G は k-maximum_degree を満たす．

$$degree\_sequence.count(\Delta(G)) > k - 1 \tag{3}$$

例えば，degree_sequence = $[5, 5, 3, 2, 2, 2, 1]$ のとき，グラフ G は 2-maximum_degree を満たす．最大次数を持つノードが少なくとも k 個あるとき，リンク予測後の匿名加工データは k-maximum_degree を満たすと言える．

3.3 概要

今回従来の手法である k-degree による匿名化とは異なる手法として，以下の 3 つの手法を提案する．

- リンク予測 + k-maximum_degree による匿名化
- リンク予測 + k-degree による匿名化
- k-degree による匿名化 + リンク予測

リンク予測には CCPA[2]，匿名化には提案手法 k-maximum_degree，および従来手法 k-degree[3] を用いる．リンク予測の際，パラメータは $\alpha = 0.8$，追加するリンクは正規化された CCPA スコアのうち，0.9 以上の値をとるリンクを予測リンクとして設定した．

4 評価

4.1 実験設定

リンク予測を用いた匿名化手法について有用性の評価を行った．今回，実験に用いるデータとして，以下の方法で 20%および 50%削除したデータの匿名加工データを用意した．

(a)k-degree を用いた匿名化
(b)リンク予測+k-maximum_degree による匿名化
(c)リンク予測+k-degree による匿名化
(d)k-degree による匿名化+リンク予測

提案手法との比較対象として，(a) のような k-degree を用いた匿名加工データを用いることとした．

4.2 データセット

実験では, 実データとして Facebook のデータセットを用いる．Facebook データは友人関係の情報が含まれており，4,039 個のノードと 88,234 本の無向エッジで構成される．

4.3 評価

実験に際して，リンクを一部削除した影響により，ノード数は 20%削除したデータのとき 4,021 個，50%削除したデータのとき 3,951 個に減少した．

実験の結果，各条件におけるリンク数および元データと一致するリンクの数は表 1，表 2 のようになった．

5 考察

実験の結果、(a) k-degree(従来手法) と比較して，3 つの提案手法 (b)(c)(d) は，元データの 88,234 本により近いリンク数の匿名加工データを生成できることが分かった．

表 1: 20%欠損するデータでの各条件におけるリンクの数

	(a)k-degree	(b)CCPA+k-maximum_degree
k=3 (一致数)	71258 (28508)	73342 (72645)
k=6 (一致数)	72284 (28528)	74409 (72669)

	(c)CCPA+k-degree	(d)k-degree+CCPA
k=3 (一致数)	73971 (28523)	71259 (28509)
k=6 (一致数)	75074 (28541)	72287 (28530)

表 2: 50%欠損するデータでの各条件におけるリンクの数

	(a)k-degree	(b)CCPA+k-maximum_degree
k=3 (一致数)	44528 (17596)	66199 (58644)
k=6 (一致数)	45114 (17585)	67330 (58678)

	(c)CCPA+k-degree	(d)k-degree+CCPA
k=3 (一致数)	66313 (17701)	44529 (17597)
k=6 (一致数)	67203 (17822)	45128 (17590)

また，提案手法の中でも (d)k-degree+CCPA は．最もリンク数の変化量が小さくなっている．その要因として k-degree を用いて匿名化する際，データの構造的な特徴を考慮されないままグラフの変更がされていることが挙げられる．そのため，(d)k-degree+CCPA では匿名加工データ内の共通の隣接と中心性の値が低くなることで CCPA スコアが低くなり，リンク予測によるリンク追加数が少なくなっている．

元データと一致するリンクの数に関しては，(a) k-degree(従来手法) と比較して，3 つの提案手法 (b)(c)(d) の一致するリンクの数が上回る結果となった．特に，提案手法 (b)CCPA+k-maximum_degree では，リンクの追加によってグラフの変更を行い，匿名加工データの最大次数を持つノードが k 個以上になるようにしている．そのため，(b)CCPA+k-maximum_degree は提案手法の中でもリンク数の一致数が最も大きく，元のグラフの特徴を残していると考えられる．

6 おわりに

本稿では，リンク予測を用いた k–匿名化手法について実際のソーシャルネットワークデータへ適応し，有用性について検討を行った．また、実データに欠損が含まれることを想定し，リンク予測を行ったデータの匿名加工時における匿名性指標についても言及した．今後は、他の実データや有用性および安全性に関する追加実験を行い，よりデータの特徴を残したリンク予測と匿名化手法について検討を行う．

謝辞

本研究は JSPS 科研費 JP21H03496，JP22K12157 及び JST，さきがけ，JPMJPR1934 の助成を受けたものです．

参考文献

[1] L. Sweeney. k–anonymity: a model for protecting privacy. Int. J. on Uncertainty Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570(2002).

[2] Ahmad, I., Akhtar, M.U., Noor, S. et al. Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm. Sci Rep 10, 364 (2020).

[3] Liu, Kun, and Evimaria Terzi. Towards identity anonymization on graphs. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 93-106(2008).