6W-03

グラフデータにおける匿名化コストの評価と検討

小早川真 †　　小林亜樹 †
† 工学院大学工学部情報通信工学科

1 はじめに

個人の特定などの危険性を取り除く技術の一つとして匿名化が注目されている．表形式データにはSweeneyが「k-anonymity」[1] を，ソーシャルネットワークデータのようなタプル同士が関係するグラフデータには，Zhou らが「k-neighbor」[2] を提案している．本研究では，ソーシャルネットワークデータに対する匿名化における情報損失の低減を目的として，既存アルゴリズムの改良手法を提案し，問題点を議論する．

2 既存研究

2.1 概要

シンプルグラフ $G = (V, E, L, \mathcal{L})$ で，V は頂点集合，E は辺集合，L は階層的な属性集合，$\mathcal{L}$ は頂点から属性を導くラベリング関数と定義し $\mathcal{L}\colon V \to L$ である．k-neighbor は，頂点の近傍（ホップ数 1 の範囲) について，同型の部分グラフが最低 k 個存在しているものと定義され，Zhou らの手法では，部分グラフに頂点を加えるために，元のグラフにはない辺を新たに加えることで匿名化を行う．本稿では，この追加する辺を擬製辺と呼ぶ．まとめると，属性の一般化と擬製辺の追加によってソーシャルネットワークデータを匿名化する．

属性 l_1 を l_2 に一般化したことによる情報損失 (Normalized Certainty Penalty) を式 (1) に示す．

$$\mathrm{NCP}(l_2) = \frac{\mathrm{size}(l_2)}{\mathrm{size}(*)} \tag{1}$$

$\mathrm{size}(l_2)$ は属性階層における l_2 の子孫の数を表しており，$\mathrm{size}(*)$ は属性階層における葉の総数を表している．頂点 u_1, u_2 を同型化した匿名化コストは式 (2) の関数として定義され，コストの低い頂点同士を同型化する指標に使われる．

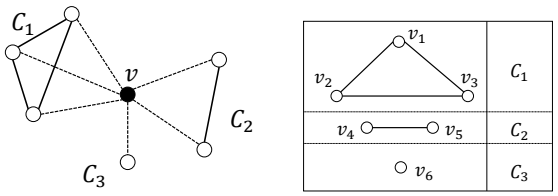
$$\begin{aligned}\mathrm{Cost}(u_1, u_2) = {} & \alpha \cdot \sum_{v' \in H'} \mathrm{NCP}(v') \\ & + \beta \cdot |\{(v_1, v_2) | (v_1, v_2) \notin E(H), (\mathcal{A}(v_1), \mathcal{A}(v_2)) \in E(H')\}| \\ & + \gamma \cdot (|V(H')| - |V(H)|)\end{aligned} \tag{2}$$

式 (2) における α, β, γ は重みパラメータである．H, H' は同型化前後の u_1, u_2 の近傍の和である．v_1, v_2 は H に含まれる頂点であり，$\mathcal{A}(v_1), \mathcal{A}(v_2)$ は H' に含まれる頂点である．(v_1, v_2) は v_1 と v_2 間の辺である．第 1 項は、属性を一般化した際の情報損失を，第 2 項は，擬製辺の総数を，第 3 項は，部分グラフに新たに加えられた頂点の総数を表している．

同型化処理の中で擬製辺と部分グラフ内に頂点を追加する必要がある場合は次に示す優先順位をもとに選択する．

1. 未匿名化頂点で最小次数を持つ頂点
2. 1. のような頂点が複数ある場合，その中でもっとも属性が類似している頂点
3. 既匿名化頂点で最小次数を持つ頂点
4. 3. のような頂点が複数ある場合，その中でもっとも属性が類似している頂点

部分グラフを同型にする処理はコンポーネントと呼ばれる誘導部分グラフを用いて行う．頂点 v を含む v の近傍が図 1(a) のような場合，頂点 v のコンポーネント集合 $\mathrm{Neighbor}_G(v)$ は v に隣接する頂点による誘導部分グラフとして定義される．これを式 (3) に示す．



(a) 頂点 v の近傍　　(b) 頂点 v のコンポーネント

図 1: 頂点 v の部分グラフとコンポーネント

$$\mathrm{Neighbor}_G(v) = \{C_1(v) + C_2(v) + C_3(v)\} \tag{3}$$

2.2 問題点

匿名化を過度に行うと，情報が大幅に失われてしまう．しかし，後に述べるように Zhou らの手法ではグラフデータに対して過度な匿名化が行われてしまうことがある．その原因は部分グラフ内に頂点を追加する際にあり，既存の優先順位では考慮されていない．グラフデータに対して過度な匿名化が行われてしまうような状況を次に示す．

- 頂点 u と v が接続されている状態でこの 2 つの頂点を同型化しようとする．その際に，頂点 v の部分グラフに頂点を追加する必要があり，追加し

Evaluating of anonymization cost in graph data
†Makoto Kobayakawa　†Aki Kobayashi
†Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

た頂点が頂点 u の部分グラフにある頂点を選択した場合

- 擬製辺を追加する際に基準頂点からホップ数2の位置にある頂点を選択した場合

入力グラフと k の値によって，このような匿名化コストが跳ね上がる条件を満たすと，匿名化できたとしても匿名化コストが高くなってしまい，元データの情報が大幅に失われてしまう．ソーシャルネットワークのようなスモールワールド性を持つグラフ構造の場合，このような条件を満たす場合が多い．

3 提案手法

本節では，2.2節で上げた問題点に着目し，部分グラフに追加する頂点の優先順位を提案する．既存手法では次数と属性のみを考慮し部分グラフに追加する頂点を選択している．一方，前述の問題点ではホップ数2の頂点の選択が共通している．その点に注意して次数，属性，ホップ数の3つを考慮した頂点選択の優先順位を次のように提案する．

1. 未匿名化且つ基準頂点からホップ数2以上の中で最小次数を持つ頂点
2. 1. のような頂点が複数ある場合，その中でもっとも属性が類似している頂点
3. 未匿名化で最小次数を持つ頂点
4. 3. のような頂点が複数ある場合，その中でもっとも属性が類似している頂点
5. 既匿名化且つ基準頂点からホップ数2以上の中で最小次数を持つ頂点
6. 5. のような頂点が複数ある場合，その中でもっとも属性が類似している頂点
7. 既匿名化で最小次数を持つ頂点
8. 7. のような頂点が複数ある場合，その中でもっとも属性が類似している頂点

この優先順位により，部分グラフに頂点を追加した際に発生する部分グラフに対する過度な加工を抑制する．

4 実験

Zhou らの手法と提案手法をスモールワールド性を持つグラフに適用した際の匿名化コストと困難さを調査するために人工データを用いて評価実験を行った．

4.1 環境と条件

実験ではPython言語上でライブラリにNetworkXを用い，スモールワールド性を持つグラフを用いた．重みパラメータ $\alpha = 0, \beta = 1.0, \gamma = 1.0$ とし，$k = 3$ とした．頂点同士の匿名化において匿名化コストが50を超えた場合にはコストが大きすぎるとして，同型化処理を中止した．

4.2 結果と考察

結果を表1に示す．既存手法では匿名化されたグループの2つ目を，提案手法では匿名化されたグループの6つ目を作成しようとした際に，匿名化コストが50を超える頂点と同型化することになったため同型化処理を中止した．既存手法は3個，提案手法は15個の頂点のみが匿名化された状態になった．

表 1: 実験結果

指標＼グラフ	匿名化前	既存	提案
頂点数	100	100	100
うち匿名化		3	15
辺数	217	221	310
平均次数	4.3	4.4	6.2
直径	10	10	7
平均パス長	4.8	4.7	3.5

提案手法は既存手法よりも多くの頂点を匿名化することができたが，完了できなかった．既存手法では匿名化できた頂点が3個にとどまり実用的でない．提案手法では匿名化前の辺数に対して，擬製辺は4割を超えグラフの構造が大きく変化した．同型化中止の原因として，頂点数の少なさ，匿名化コストが跳ね上がるような条件を満たしたためと考える．スモールワールド性により平均パス長が短いグラフは部分グラフに追加する頂点が基準頂点から近い位置にあることが多く，過度な匿名化が起こりやすい．また，擬製辺を追加する手法では擬製辺の追加によってその後の匿名化コストが増えていくこともある．

スモールワールド性を持つグラフに対して，Zhou らの手法から改良できたといえるが，擬製辺を追加するだけでは実用は難しいのではないかと考察する．

5 おわりに

本稿では，k-neighbor 手法の現実的なグラフデータへの適用において生じる問題点について論じ，改良手法を提案した．大規模な実験や次数分布，匿名化困難さや情報損失などの定量評価は今後の課題である．

参考文献

[1] L.Sweeney. “k-anonymity:a model for protecting privacy,” International on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp.557-570, 2002.

[2] Bin Zhou and Jian Pei. “Preserving privacy in social networks against neighborhood attacks,” In Proc. ICDE’08, pp.506-515, 2008.