5N-5

ソーシャルネットワークにおける距離関係を考慮したk-匿名化

岡田莉奈† 渡辺知恵美‡ 北川博之‡

†筑波大学情報学群情報科学類 ‡筑波大学システム情報系情報工学域

1 はじめに

ソーシャルネットワークデータ (SN データ) を研究やデータ分析の目的で一般的に公開するためには，利用者のプライバシー保護のための匿名化が必要である．SN データはユーザをノード，ユーザ間の関係をエッジで表現したグラフ構造とすることが一般的であり，匿名化をする際にはグラフ構造を考慮する必要がある．SN データ匿名化の一既存手法として任意のノードの隣接ノードから成るサブグラフに着目し，同型のサブグラフが少なくとも k 個存在するようにノイズエッジを追加する k-neighbor という匿名化がある．しかし，k-neighbor を実現するアルゴリズムではノイズエッジの追加によりノード間の距離関係が大きく変化する．そこで，本研究では距離関係を最小限に抑える k-匿名化のアルゴリズムを提案する．

2 k-neighbor

2.1 k-neighbor の匿名化指標

k-neighbor では，SN データを重複するエッジのないシンプルな無向グラフ $G = (V, E, L, \mathcal{L})$ で表し，このグラフデータを匿名化するための指標について定義している．グラフ G において各ユーザのプロパティはノードのラベルとして簡潔に表現されている．V はノードの集合，$E \subseteq V \times V$ はエッジの集合，L はラベルの集合，$\mathcal{L} : V \to L$ は各ノードからラベルを結びつけるラベル関数である．$V(G)$, $E(G)$, L_G, $\mathcal{L}_G$ は，それぞれグラフ G のノード集合，エッジ集合，ラベル集合，ラベル関数のことである．

攻撃者は，攻撃対象となるユーザのラベルとそのユーザの友人のネットワークトポロジーに関する背景知識を持ち，攻撃対象ユーザのノードを特定したいと想定する．これは攻撃者が対象ノードの隣接ノードからなるグラフを知っており，そのグラフにマッチするグラフ G 内のサブグラフを探すことと対応付けられる．例として図 1 を考える．図 1(a) は SN データであり，図 1(c) は攻撃者の持つ背景知識であるサブグラフである．(a) 内には (c) と同様のサブグラフが 1ヶ所しかないので，Ada が特定されてしまう．しかし，図 1(b) のように Harry と Irene の間にノイズエッジを加えることによって，(b) 内には (c) と同様のサブグラフが 3ヶ所存在することになる．このように，k-neighbor では SN データ内の任意のノードの隣接ノードから成るサブグラフが少なくとも他に $k-1$ 個存在することを匿名化指標としている．

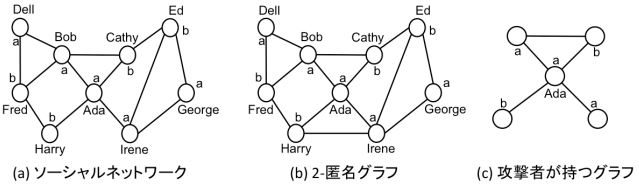

図 1: k-neighbor の例

2.2 k-neighbor を実現するアルゴリズム

文献 [1] では，k-neighbor の定義に加えて，k-neighbor を実現するアルゴリズムを提案している．まずノード v の誘導部分グラフを隣接コンポーネント集合 $Neighbor_G(v)$ として表し，ある 2 ノード u, v 間の $Neighbor_G(u)$,$Neighbor_G(v)$ を匿名化するためのコスト $Cost(u, v)$ を定義している.. 例えば図 2(a) における $Neighbor_G(v_1)$ は C_1, C_2 となる．匿名化コストが小さな k 個のノードをグラフ G から見つけて同型にすることによって k-neighbor を保証する．2 つのノード u, v の隣接コンポーネント集合 $Neighbor_G(u)$,$Neighbor_G(v)$ の匿名化コストは，互いのコンポーネント集合を同型にするためにノードのラベルの一般化コスト，追加するエッジ数，コンポーネントに新たに含めるノード数によって求められる．

2.3 既存アルゴリズムの問題点

文献 [1] のアルゴリズムでは，隣接コンポーネントを同型化する処理中でコンポーネント内に新たに含める

The k-neighbor anonymization of social network data with keeping distances between nodes

Rina OKADA†(okarina@kde.cs.tsukuba.ac.jp),
Chiemi WATANABE‡(chiemi@cs.tsukuba.ac.jp) and
Hiroyuki KITAGAWA‡(kitagawa@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba
‡Faculty of Engineering, Information and Systems, University of Tsukuba
§Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency

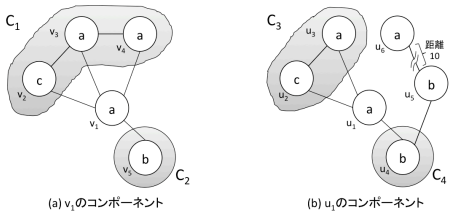

図 2: コンポーネントの例

ノードを選択する際に，グラフのユーティリティに大きく影響を与えることがある．SN データの特徴であるスケールフリー性とスモールワールド性を保つために (1) 最少次数を持つノード，(2) 最少次数を持つノードが複数ある場合は最もラベルが類似ししているノード，という優先順位でノードを選択する．

しかしながら，この優先順位ではグラフ G 内のノード間の距離関係が大きく変化してしまう可能性がある．例えば，図 2 の場合，ノード u_3 とノード u_6 の間にノイズエッジを追加すると C_1 と C_3 が同型になるが，同型化前のノード u_3 とノード u_6 の距離は 13 であり，同型化後の距離は 1 である．このように匿名化の前後でノード間の距離関係が大きく変化することがある．そこで．我々はユーティリティとして，スケールフリー性とスモールワールド性のプロパティに加えて，ノード間の距離関係の変化も考慮する改良アルゴリズムを提案する．

3 提案アルゴリズム

本研究では，既存のアルゴリズムを SN データ内のノード間の距離関係の変化を最小限に抑えるように改良した．これを Algorithm1 に示す．このアルゴリズムでは，接続元のノードに近い次数の低いノードから順に幅優先探索を行い，片方のコンポーネントと同じラベルを持つノードを選択する．また，ラベルが一致するノードがない場合は一般化したラベルを Algorithm1 の入力にして再探索を行なう．はじめからラベルの一般化を行わない理由は，一般化をたくさん行なうと一般化が伝搬し，情報損失が大きくなるからである．

4 関連研究

k-neighbor とは異なる指標として，次数が同じノードが少なくとも k 個あることを保証する k-degree 匿名化 [2] がある．しかし，この匿名化指標ではラベルを一切考慮していないため攻撃のリスクが高いと考えられる．さらに，k-degree を l-多様性まで考慮しているものが k-degree-l-diversity 匿名化 [3] である．匿名化指標では，ノイズエッジを加える際にノード間の距離の変化が大きくならないようなノイズエッジの加え方を議論しているが，どのようにしても距離の変化が大きくなってしまう場合はノイズノードを加えることによってデータ分析の精度が低減しないように工夫している．本研究でも今後は距離の変化が低減できない場合はノイズノードを加える拡張を行なうことも視野にいれている．

Algorithm 1：距離関係を考慮した component に追加するノードの選択

Input:
接続元のノード v，接続先のノードのラベル *label*;

1: **ノード v の隣接ノード $s(\notin component)$ の次数が昇順になるようソートし，待ち行列 Q に入れる (ただし，ノード s はまだ匿名化されていないノード);**
2: **while** $|Q| > 0$ **do**
3: **待ち行列 Q から s を取り除く；**
4: **if** $s.label() = label$ **then**
5: **return s；**
6: **end if**
7: **ノード s の隣接ノード $s'(\notin component)$ の次数が昇順になるようソートし，待ち行列 Q に入れる (ただし，ノード s' はまだ匿名化されていないノード)；**
8: **end while**

5 まとめと今後の予定

本論文では，ノード間の距離関係を考慮した k-neighbor を実現するアルゴリズムを提案した．今後は実データによる評価を行う予定である．

参考文献

[1] Bin Zhou 0002 and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowl. Inf. Syst.*, 28(1):47–77, 2011.

[2] Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *In Proceedings of ACM SIGMOD*, pages 93–106, 2008.

[3] Mingxuan Yuan, Lei Chen, Philip S. Yu, and Ting Yu. Protecting sensitive labels in social network data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):633–647, 2013.