科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　５　日現在

機関番号：　１２１０２
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７０００８８

研究課題名（和文）プライバシ保護に配慮したソーシャルネットワーク分析環境の開発

研究課題名（英文）Development of Social Network Analysis Which Considers Privacy Preservation

研究代表者
渡辺　知恵美（Watanabe, Chiemi）

筑波大学・システム情報系・助教

研究者番号：２０３６２８３２

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：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．本研究ではグラフ匿名化手法をサーベイしたうえで，プライバシ保護を考慮したSNデータ分析を実現するための匿名化アルゴリズムの提案を行った。特に我々はk-degreeおよびk-automorphismという二つの匿名化指標に着目し，属性付きノードでグラフの特徴を維持できる匿名化アルゴリズムを提案し実装した。

研究成果の概要（英文）：In order to publish the social network data to provide the data to many data analysts, we should apply anonymization for protecting users' privacy in the social network service. We investigate the useful and secure anonymization algorithm for social network data. Especially we focus on two anonymization measure named k-degree and k-automorphism, and we propose the algorithms which can deal with the social network graph with node attributes and the algorithms. In addition we propose the algorithms which can keep the feature of original SN data after anonymization.

研究分野：　データベース

科研費の分科・細目：　メディア情報学・データベース

キーワード：　プライバシ保護　匿名化　ソーシャルネットワーク

１．研究開始当初の背景

近年，ソーシャルネットワークサービス(SNS)が大変注目を集めている．SNS における利用者のコメントや行動履歴は膨大な量となり，SNS に集められた膨大なデータを収集して分析し，マーケティング戦略や社会科学分析，SNS に関する新たな技術開発に生かしたいという需要が非常に高まっている．しかしながら，そのようなデータを第三者に提供する場合，データには個人情報が多く含まれているため，個人情報の漏えいを防ぐことを保証しつつデータを提供しなければならない．個人情報の漏えいを防ぐには，氏名や住所，電話番号など個人を一意に特定される情報を削除するだけでは十分ではないことが知られている．性別や年齢，SNS から得られる友人関係や行動履歴など，本来分析に使われるデータそのものに関しても，それらの値の組合せから個人情報が推測されないように十分な匿名化を行わなければならない．またそれらの匿名化によってデータの値の一般化およびノイズの付与，一部のデータの改ざんなどが行われた場合，それらの編集を踏まえたうえでの分析が必要であり，通常のマイニングツールを適用した場合不正確な分析結果を招いてしまう可能性がある．

２．研究の目的

我々は SNS 会社が一般に公開されていないデータを研究や分析目的で第三者に提供する場合を想定した，個人情報を十分に匿名化したデータ公開法と匿名化データを用いたデータマイニングツールを提供するソーシャルネットワークデータ分析環境を開発する．

一般に SNS によって収集されるデータはグラフ構造で表現される．例えばユーザ同士の参照関係や友人関係は，ユーザをノード，関係をエッジで表わしたグラフで表わされる．またユーザとコミュニティ（趣味やイベントなど）の参加関係はユーザノードとコミュニティノードとの二部グラフで表わすことができる．また各ノードやエッジにユーザの情報（年齢や所在地，趣味など）をプロファイル情報として表すことでプロファイルとユーザ間の関係を組み合わせた分析（「30 代の女性のレビュー閲覧と口コミの広がりの関係」など）を行うことができる．

本研究では既存の SNS データ匿名化のサーベイを行った上で、現状の匿名化に関して以下の点を解決するアルゴリズムを提案した。

(1) 既存のアルゴリズムでは匿名化時にグラフの構造的特性のみに着目し、各ノードの属性を踏まえた匿名化手法が未熟であるという問題点に対し、ノードの属性を踏まえた匿名化手法を提案した。

(2) 匿名化後の SNS データが元のデータの特徴を大きく損なうことによる問題を解決するために、グラフ特性の一つである距離関係の変化を抑制するアルゴリズムを提案した。

(3) 利用者が個々に情報を匿名化するタイプの SNS サービスにおいて，利用者が意図しない情報漏えいを検出するためのフレームワークを提案した。

３．研究の方法

SNS 匿名化で使用できるグラフ匿名化指標には k-degree 匿名化，k-neighbor 匿名化および k-isomorphism 匿名化がよく知られている。k-degree 匿名化は各ノードの次数に着目し，同じ次数のノードが k 個以上あるように匿名化をする指標である。k-neighbor 匿名化は各ノードの隣接ノードで構成される部分グラフに着目し，部分グラフの構造でグループ化をしたときすべてのグループが k 個以上の部分グラフで構成されるように匿名化する指標である。k-automorphism は隣接ノードに限らず各ノードの n-hop 先のノードまで含めた部分グラフまで考慮しても、同じ構造の部分グラフが k 個以上存在するという匿名化指標である。

(1) ノードラベルを考慮した k-automorphism 匿名化アルゴリズムの提案

上記に挙げた匿名化指標のうち k-degree および k-neighbor はノードの属性をノードラベルに置き換え、ノードラベル情報も含めて匿名化すること想定している。その一方 k-automorphism では既存のアルゴリズムではノード属性を考慮していない。これらのことから，研究の 1 点目として k-automorphism における既存の匿名化アルゴリズムを改良し，ノードラベルを踏まえたアルゴリズムを提案した。

k-automorphism アルゴリズムでは、任意のホップ数の部分グラフに対して、同じ構造の部分グラフが k 個以上存在するように匿名化処理を施す。そのためのアプローチとして、類似した部分グラフを相互に接続し自己相似化することによって、任意のホップ数でも同じ構造の部分グラフが同型になるようにノイズノードを追加している。例えば図 1 左に示す二つの部分グラフに対して，v2-v7, v6-v8 の辺を加えることにより，二つのグラフは自己相似な部分グラフとなる。

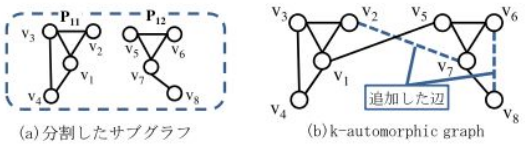
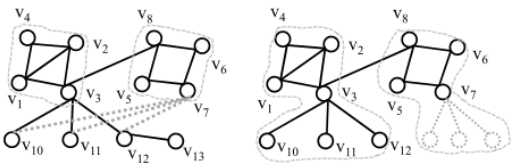


図 1: k-automorphism による匿名化

ただし既存のアルゴリズムではノードの属性を考慮していないため、これを考慮するよう拡張した。まずノードの属性を類似度によってグループ化し，同じグループのノードはすべて同じ属性値になるように一般化する．同じグループのノードにはグループ番号によるラベルを付けることによって、ラベルつきグラフの状態にする．グループ数は k とは異なる任意の数をパラメタとして設定する。また自己相似グラフのコスト関数を以下のように定義し、既存の手法の問題点であった自己相似化によるノイズエッジの過剰な追加を抑える。

$$COST(U_i) = AlCost(U_i) + CrossEdgeCost(U_i) + OutEdge(u_i) + ExpandCost(U_i)$$



(a)OutEdge Cost　　(b)Expand Cost

図２：ノードラベルつき k-automorphism のコスト関数

これにより、匿名化による辺の追加を抑えつつグラフの属性を考慮した匿名化アルゴリズムを実現した。

(2) ノード間の距離関係を抑制する匿名化アルゴリズムの提案

(1)で取り上げた k-automorphism 匿名化とは異なり，k-degree 匿名化及び k-neighbor 匿名化はノードラベルも考慮した指標である。しかしながら、これらの匿名化指標を実現するためのアルゴリズムには，元のグラフの特性を損なってしまう問題点が残っている。我々はその中で k-neighbor 匿名化アルゴリズムに着目し，データを匿名化することによって元のデータの特徴を抑制する手法を提案した。k-neighbor では，攻撃者が攻撃対象となるユーザの友人関係を知っていることを想定し，隣接ノードから成るサブグラフに着目する。そして，任意のノードが少なくとも他の $k-1$ 個のノードとサブグラフが同型であれば，k-neighbor 匿名であると定義している。しかしながら k-neighbor 匿名化を実現するアルゴリズムでは、辺の追加によってグラフの特徴の一つであるノード間の距離関係が大きく変化することを本研究にて指摘し，ノード上の距離関係をできるだけ維持することのできるアルゴリズムを提案しその検証を行った。

k-neighbor 匿名化を実現する既存のアルゴリズムでは，各ノードの隣接ノードによるサブグラフを強連結成分によるサブグラフ（隣接ノードコンポーネントと呼ぶ）集合に分割し，隣接ノードコンポーネント集合が類似した k 個以上のグループを作る。そのグループにおいて，ノードコンポーネントのグラフ構造が同型となるようノイズエッジを追加する。同型化の際、隣接ノードコンポーネントのノード数が異なる場合は，ノード数が少ないコンポーネントに対してノードを追加する必要がある。このノードの選択方法として、既存の手法は特に基準が決められておらず、次数の小さなノードが選択されていた。図 3(a)の例で示される例の場合、C3 のコンポーネントに u6 を追加すると，距離が 10 以上あったノード間の関係が一気に隣接ノードとなり，周辺のノードも含めて距離関係が大きく崩れてしまう問題がある。

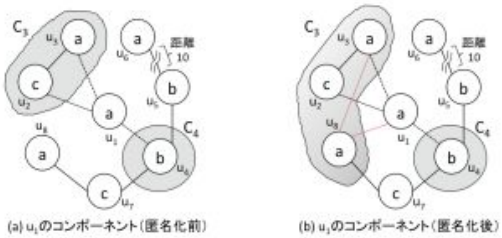


図３：提案手法による k-neighbor 匿名化手法適用の例

提案手法では，コンポーネントを追加するアルゴリズムを提案した。ダイクストラ法をベースに，コンポーネントに含まれるノードから距離が近いノードを対象に接続可能なノードを探索する。接続可能な条件は既存の手法では十分に記述されていなかったため、我々の調査によって接続可能条件を定義し、それに当てはまる、コンポーネントから最も近いノードを選択する。アルゴリズムは以下のとおりである。これによって、匿名化前のグラフの距離関係の変化をできるだけ抑制することができる。図３(b)では本提案アルゴリズムを適用した結果、u8 のノードを選択することになる。

(3) 利用者が意図しない情報漏えいを検出するためのフレームワーク

個々の SNS データが匿名化されていても、複数の SNS データのマッチングをとることによって脱匿名化される可能性もある。我々はこのようなケースに対応するため、複数の SNS を利用しているユーザが個々の SNS で匿名化をしていた場合でも、複数の SNS により脱匿名化に至ってしまうプライバシリスクを求めるフレームワークを提案した。これまでの 2 項目とは異なり，この研究では利用者個々人が自らの責任で匿名化を行っていることを想定している。（匿名化を行わなくても良い。）我々はこのプライバシリスクをアカウント到達可能性と定義し、アカウント到達可能性を計算するためのフレームワークを定義した。

アカウント到達可能性（Account Reachability）とは攻撃者が利用者の既知のアカウントから別のアカウントを見つけ出す可能性を表す．たとえば，ある利用者が二つの異なる SNS のアカウント s1，s2 をそれぞれ持っているとする．また攻撃者は利用者の SNS アカウントのうち，s1 のみしか知らないとする．攻撃者は s1 の情報をもとにして，まだ知らないアカウントである s2 をさまざまな手法を通して見つけ出そうとする．ここで攻撃者は s1 のプロファイルや投稿内容から s1 のキーワードを抽出し，検索エンジンなどを用いて検索を行い，s2 になりうるアカウントの候補を取得する手法をとったとする．このとき，取得した候補アカウントそれぞれと s1 から取得したキーワードをもとに s1 との類似度をはかり，s2 が s1 のアカウントであると特定していく．アカウント到達可能性 AR(s1->s2) の計算式を以下に示す．

$$AR(s_1 \to s_2) = max_{q \subseteq Q}(AR(s_1, s_2, q))$$

$$Q = GenQueries(s_1.prof, s_1.msg).$$

$$AR(s_1 \to s_2, q) =$$

$$Match(s_2, Cand(q)) * \frac{Score(s_1, s_2)}{\sum_{c \in Cand(q)} Score(s_1, c)}$$

$$Match(s_2) = \begin{cases} 1 \; if \; s_2 \in Cand(q) \\ 0 \; else \end{cases}.$$

４．研究成果

３．にて述べた 3 つの提案手法の実験結果を示す。

(1) ノード属性を考慮した k-automorphism 匿名化

本提案手法を prefuse graph という仮想の SNS グラフ（ノード数 129，エッジ数 161）に適用し検証した。k の値が低い場合でも 100 本以上のエッジが追加されており，

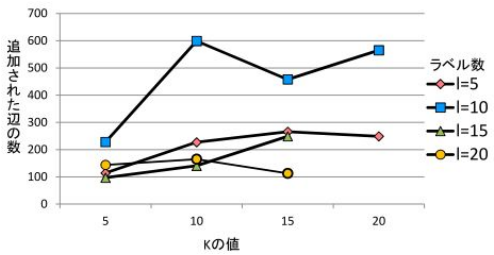


図４：追加されたエッジ数

本匿名化指標を満たすには非常に多くのエッジが追加されてしまうことが分かる。ただし､匿名化後のグラフを k-means クラスタリングし、匿名化前の結果との相関をもとめたところ、クラスタリング相関が 0.5～0.8 にとどまっており､匿名化による劇的な変化は抑えられているとわかった。

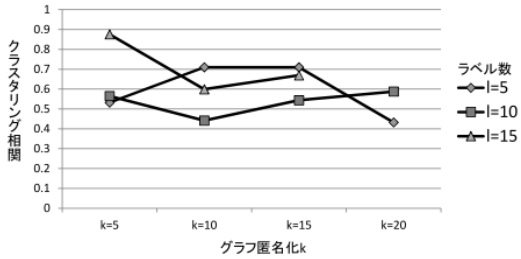


図５：匿名化前と匿名化毎のクラスタリング結果の相関

(2) 距離の変化を抑制する k-neighbor 匿名化アルゴリズム

Python のライブラリである networkX を用いてスモールワールド性を持った人口データ（ノード数 300，エッジ数 600，ラベル数 3）を作成し検証を行った。図 6 の縦軸は匿名化による距離変化平均を示している。既存手法と比較して多少であるが距離変化が小さく抑えられていることが分かる。嵯峨あまり大きくないのはもともとスモールワールド性があるグラフであるため、距離の大きなノードが多くないためである。

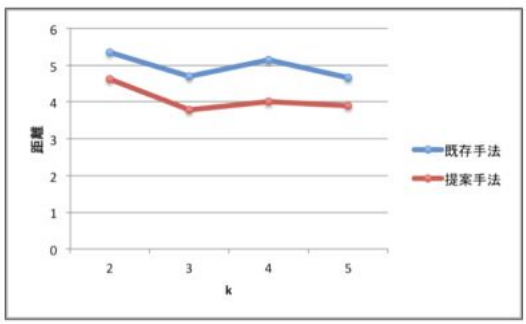


図６：匿名化後による距離変化平均

(3) 利用者が意図しない情報漏えいを検出するためのフレームワーク

我々は Twitter と Facebook のアカウントを所有する 50 名を対象に、Facebook から Twitter へのアカウント到達可能性を計算し検証した。50 名のうち 26 名がアカウントが到達化されても構わない（NotCare）で，それ以外が到達されたくない（NotWant）である。図 7 は計算結果である。多くの利用者はユーザの意図する匿名化が実現できているが、数名到達されたくないにもかかわらず到達可能性が高くなっているユーザがいる。このようなユーザがこの結果を見てプライバシリスクを自覚することができる。

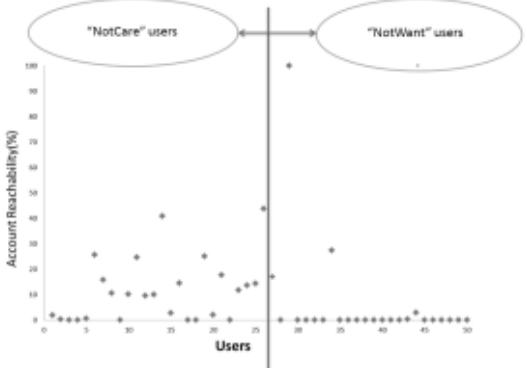


図７：50 名のユーザによる Facebook から Twitter へのアカウント到達可能性

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 5 件）

Ayano Yoshikuni and Chiemi Watatanabe: Account Reachability: A Measure of Privacy Risk for Exposure the User's Multiple SNS Accounts, the 15th International Conference on Information Integration and Web-based Applications and Services (iiWAS2013) (2013), 2013 年 12 月 2 日～4 日, ウィーン, オーストリア

岡田莉奈，渡辺知恵美，北川博之：ソーシャルネットワークデータの距離関係の変化を抑制する k 匿名化アルゴリズム，第 6 回データ工学と情報マネジメントに関するフォーラム, B2-2(2014)，2014 年 3 月 2 日～4 日，兵庫県淡路島

６．研究組織
(1)研究代表者
渡辺 知恵美（Watanabe Chiemi）
筑波大学 システム情報系　助教
研究者番号：20362832