共通識別子を持つレコード群のデータ匿名化手法

髙橋翼† 側高幸治† 豊田由起† 竹之内隆夫† 森拓也† 興梠貴英‡

† 日本電気株式会社 情報・ナレッジ研究所
211-8666 神奈川県川崎市中原区下沼部 1753
t-takahashi@nk.jp.nec.com

‡ 東京大学大学院医学系研究科 22 世紀医療センター
113-8655 東京都文京区本郷 7-3-1

あらまし 蓄積されたパーソナル情報の活用にはプライバシ保護が必要であり，保護手法の一つとして k-匿名化による匿名性保証技術が知られている．従来の k-匿名化方式では，データ主体とレコードが一対一に対応する環境において，レコード単位の匿名性保証が検討されてきた．しかし，実際には一つのデータ主体に複数のレコードが存在する．従来の k-匿名化されたレコードは，共通の識別子によって他のレコードと突合された際に匿名性が破綻する場合がある．本稿では，データ主体に複数のレコードが存在する環境において，匿名性破綻を解消する匿名化方式を提案する．提案方式を用いた評価では，小さな情報損失で匿名性破綻を解消できることがわかった．

A Data Anonymization for Multiple Personal Records with the Same Identifiers

Tsubasa Takahashi† Koji Sobataka† Yuki Toyoda† Takao Takenouchi† Takuya Mori† Takahide Kohro‡

†Knowledge Discovery Research Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, 211-8666, JAPAN
t-takahashi@nk.jp.nec.com

‡Department of Translational Research for Healthcare and Clinical Science, University of Tokyo

Abstract Personal Records are desired to use various purposes, but it is necessary to protect privacy for data owners by data anonymization such as k-anonymization. Existing k-anonymization methods assume that a data owner has only a personal record and ensure k-anonymity for a record. However, actually, each data owner has multiple personal records with having the same identifier. In this case, k-anonymized records may break k-anonymity because the number of candidates for a data owner can narrow down less than k by linking records having the same identifier. This paper proposes a data anonymization method which prevents such record linking.

1 はじめに

診療履歴やサイト訪問履歴といったパーソナル情報が，サービスを受ける度に蓄積されている．近年，ビッグデータ活用のニーズが高まり，これらの蓄積されたパーソナル情報を他のサービスや事業に利用する二次活用の期待も高まっている．しかしながら，パーソナル情報はデータ主体の個人に関する機微な情報が記録されて

いることが多く，二次活用にはデータ主体のプライバシへの配慮が必要となる．特に診療履歴やレセプトのような医療情報は，機微性が高く，積極的な二次活用が行われていない．

一方，データ主体のプライバシを保護するために，パーソナル情報の個人を特定し得る属性を加工する技術であるデータ匿名化が研究されている．データ匿名化の中でも広く知られている k-匿名化 [1] は，個人を特定し得る属性である準識別子を加工し，同一の準識別子の組合せを持つレコードが k 個以上になるような加工を施す．k-匿名化の加工処理では，汎化 (Generalization) や削除 (Suppression) といった手法が用いられる．

実際のパーソナル情報には，ある個人に関するレコードが複数記録されていることが多い．例えば，診療履歴であれば，診察を受ける度に新たなレコードが，患者 ID などの共通の識別子 (共通識別子) を付加されて記録される．共通識別子を用いて同一データ主体 (患者) のレコードを紐づけることによって，病状の変化や傾向といった系列分析を行うことができる．

従来の k-匿名化では，共通識別子は削除される．共通識別子が付与されている場合，複数のレコードの組合せによって，個人特定性が高まる場合がある．本稿では，共通識別子を付与したままパーソナル情報のデータセットを匿名化を扱う．共通識別子を持つレコードが複数存在するデータセットに対して，従来の k-匿名化を行うと，各レコードの準識別子がそれぞれ異なる値へと加工される場合がある (図 1)．

一方，準識別子には生年や性別のようにデータ主体に対して常に不変な属性値が存在する．この不変な準識別子に対して共通の識別子を持つ複数のレコードを比較すると，準識別子の値を詳細化でき，それぞれのレコードに保証されていた k-匿名性が破られる (匿名性破綻)．

本稿では，共通の識別子を持つ複数のレコードの突合に対して頑健な匿名化を扱う．特に，これまでの k-匿名化手法を拡張し，k-匿名化されたデータセットに対して再匿名化処理を施すことで，上述の匿名性破綻を解消する．加えて，k-匿名性を拡張した k^*-匿名性を提案し，少ない情報損失で匿名性破綻を解消できる再匿名化手法を提案する．

本稿の以降の構成は以下の通りである．2 章では，本稿が扱う問題を定義する．3 章では，2 章で扱った匿名性破綻を解消するための再匿名化手法を提案する．4 章にて，提案方式と従来方式の差異を評価実験にて示す．5 章では関連研究を紹介し，最後に 6 章にて，本稿の結論を述べる．

2 問題定義

2.1 不変準識別子と可変準識別子

パーソナル情報には，個人を一意に特定する明示識別子 (Explicit Identifier)，性別や年齢などの個人を特徴付け，1 つ以上の属性値の組み合わせから個人を特定し得る準識別子 (Quasi Identifier)，病名などの他人に知られたくない機微な情報であるセンシティブ属性 (Sensitive Attribute) が含まれる．

また，本稿では準識別子を 2 つの種類に分類する．一つが，性別や生年月日のようにデータ主体に対して不変な値を持つ準識別子である．もう一つが，診療年月や購買日のようにアクティビティに対して付与され，レコードごとに異なる値を持つ準識別子である．本稿では，前者を不変準識別子 (Immutable Quasi Identifier)，後者を可変準識別子 (Mutable Quasi Identifier) と呼ぶ．

2.2 k-匿名化

Sweeney は，同一の準識別子の組み合わせを持つレコードが k 個以上存在するというレコード識別の困難さを表す k-匿名性を提案した [1]．データセットに対して k-匿名性を充足させるための加工を k-匿名化と呼び，属性値の汎化や削除などが用いられる．

既存の k-匿名化手法では，一般的に氏名や SSN といった明示識別子を削除し，仮 ID などの明示識別子と同等の機能を有する識別子の付与も行われない．しかし，実際のデータセットには，一データ主体に複数のレコードが存在する．

本稿では，共通識別子を付与されているレコード群を，同一データ主体のレコード群とする．

本稿における共通識別子は，氏名やSSID，患者IDなど明示識別子と同じように，レコード群を結合可能なランダムに再生成された識別子(代替識別子)を想定する．

ここで，元のデータセットをT, 代替識別子を付与したデータセットをT'とする．データ主体に複数のレコードが存在するとき，k-匿名性を以下のように拡張する．

定義 1 *(拡張したk-匿名性)* あるデータ主体のレコードと同じ準識別子の組合せを持つ，他のデータ主体が少なくとも*k-1*個存在する．

以降，本稿では拡張したk-匿名性をk-匿名性とする．

ここで，匿名化対象のデータセットT'とT'を匿名化したデータセットT^*を定義する．

T'は，d次元の準識別子の集合$QI = \{QI_1, QI_2, \ldots, QI_d\}$を持つ．各準識別子$QI_i$の定義域は$D_i = \{v_1, v_2, \ldots, v_d\}$とする．共通識別子$u$，レコード識別子$r$のレコードの準識別子$QI_i$の値を，$qi_i^u[r]$とする．$qi_i^u[r]$は$D_i$のいづれか一つの値を持つ$(qi_i^u[r] \in D_i)$.

匿名化されたデータセットT^*は，元のデータセットTと同じ準識別子の集合QIを持つ．匿名化(汎化)された共通識別子u，レコード識別子rのレコードの不変準識別子QI_iの値を，$qi_i^{*u}[r]$とする．$qi_i^{*u}[r]$は$v \in D_i$の集合値である$(qi_i^{*u}[r] = \{v \in D_i\})$.

上述のk-匿名性を充足させたデータセットT^*の例$(k = 3)$を図1に示す．図1はデータ主体Aの「生年」に関して$qi^{*A}_{\text{生年}}[A_1] = \{1970, 1971, 1972, 1973, 1974\}$と，$qi^{*A}_{\text{生年}}[A_2] = \{1974, 1975, 1976\}$を持つ．

2.3 共通識別子による複数レコードの突合

T^*には，一データ主体に複数のレコードが存在するため，一データ主体に一つ以上の不変準識別子が存在する．攻撃者が特定のデータ主体に関して真の準識別子の値を知っていたとしても，k-匿名性が保証されているため，攻撃対象

	代替ID	生年	性別	診療月	病名
A_1	A	1970-1974	Any	4月	糖尿病
B_1	B	1970-1974	Any	4月	高血圧症
C_1	C	1970-1974	Any	4月	糖尿病
A_2	A	1974-1976	Any	5月	緑内障
D_1	D	1974-1976	Any	5月	結膜炎
E_1	E	1974-1976	Any	5月	網膜剥離
D_2	D	1975-1978	女性	6月	結膜炎
E_2	E	1975-1978	女性	6月	網膜剥離
F_1	F	1975-1978	女性	6月	緑内障

図 1: k-匿名化データセット (k=3)

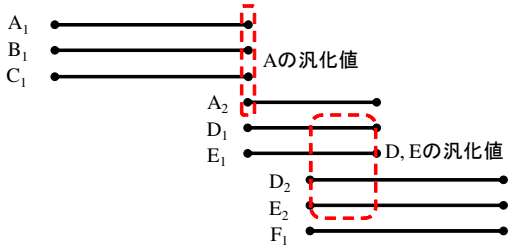

図 2: 共通識別子も持つk-匿名化レコード群の匿名性破綻

のレコードをk未満に絞り込むことができない．しかしながら，攻撃者が「同一データ主体の不変準識別子の値は同じ値を持つ」という知識を用い，共通識別子によってレコード群を突合し，不変準識別子を比較することで，当該データ主体の不変準識別子の値を詳細化することが可能である．

ここで詳細化とは，k-匿名化されたデータ主体uの準識別子に対して，uの真の準識別子の値の候補を，k-匿名化されたレコードの準識別子の値の種類(範囲)よりも少ない種類(狭い範囲)に絞り込むことである．

データ主体uに関してmレコード存在し，以下の条件が満たされるとき，少なくとも一つのレコードの不変準識別子QI_iの値が詳細化される．

$$\cap_{j=1}^{m} qi_i^{*u}[r_j] < \max_{j \in n}\{|qi_i^{*u}[r_j]|\} \qquad (1)$$

ここで，r_jはレコードの識別子であり，図1では，A_1やD_2などが該当する．

詳細化が生じる例を図 1 と図 2 を用いて示す．データ主体 A の「生年」に関してレコード A_1 と A_2 を突合すると以下のように詳細化される．

$$qi^{*A}_{\text{生年}}[A_1] = \{1970, 1971, 1972, 1973, 1974\}$$
$$qi^{*A}_{\text{生年}}[A_2] = \{1974, 1975, 1976\}$$
$$qi^{*A}_{\text{生年}}[A_1] \cap qi^{*A}_{\text{生年}}[A_2] = \{1974\} < |A_2|$$

また，データ主体 D の「生年」に関して D_1 と D_2 を突合すると以下のように詳細化される．

$$qi^{*D}_{\text{生年}}[D_1] = \{1974, 1975, 1976\}$$
$$qi^{*D}_{\text{生年}}[D_2] = \{1975, 1976, 1977, 1978\}$$
$$qi^{*D}_{\text{生年}}[D_1] \cap qi^{*D}_{\text{生年}}[D_2] = \{1975, 1976\} < |D_1|$$

A, D に対する突合は共に式 (1) を満たし，データ主体 A の生年は「1974」, D の生年は「1975-1976」へとそれぞれ詳細化された．

この詳細化によって,「同一の準識別子を持つレコードが k 個以上存在する」という k-匿名性が破られる可能性がある．

図 1 では，上述の A, D に対する詳細化によって，全てのレコードに対して同一の生年の値と考えられるレコードが $k = 3$ 未満となり，k-匿名性 (k=3) が破られる．また，あるデータ主体に対して，データ主体 A の真の生年の値 (1974) を知っていた場合，図 1 の T^* では，データ主体 A とレコード A_2 を一意に対応付けることができる．

このようにして，共通識別子を用いて複数のレコードを突合することで，特定のデータ主体の不変準識別子の値が詳細化され，匿名性破綻が生じる．

2.4 匿名性保証

本稿では前述の匿名性破綻の解消を扱う．匿名性破綻を解消するためには，共通識別子を持つすべてのレコードが同一の不変識別子の値を持つ必要がある．匿名性破綻を解消する方法には以下の 2 つが考えられる

- 匿名性破綻が生じないように k-匿名化する
- k-匿名化したデータセットを再加工し匿名性破綻を解消する

前者の方法は，全域的再符号化 (Full-Domain Generalization) を用いることで実現できる．全域的再符号化は，最もシンプルな匿名化手法であるが，それ故に情報損失量も大きい．TDS(Top Down Specialization)[2] が一例である．

後者の方法では，k-匿名化後のデータセットの同一データ主体のレコード群に対して，匿名性破綻を解消するように再加工を行う．この手法の例である局所的再符号化 (Local Recoding)[3] や多次元大域的再符号化 (Multi-Dimensional Global Recoding)[4] による k-匿名化では，前者の方法に比べて小さな情報損失を持った匿名化データを生成可能であるが，匿名性破綻が生じる可能性がある．匿名性破綻が生じた場合，再加工により情報損失が増大してしまう．

本稿では後者の方法を取り上げ，再加工による情報損失を低減した手法を提案する．以降，匿名性破綻を解消するための再加工を再匿名化と呼ぶ．

3 再匿名化

3.1 複数レコードの突合に対する再符号化

まず，匿名性の破綻を解消するために，共通識別子を持つレコードの不変準識別子が詳細化されないようにそれぞれ加工を行う．ここでは，共通識別子を持つすべてのレコードの不変準識別子を同じ値に汎化する．汎化による情報損失を最小限に抑えるために，同一データ主体のレコード群に対して，すべての不変準識別子の組合せを包含する最小の超集合へと加工し，同一データ主体の全ての不変準識別子を qi^{**u}_i とする．

$$qi^{**u}_i = \cup_{j=1}^{n} qi^{*u}_i[r_j] \tag{2}$$

例えば，データ主体 A のレコード群の生年には，A_1 に「1970–1974」，A_2 に「1974–1976」が存在する．このとき，これらを包含する最小の超集合を新たな値とする．具体的には,「1970–1976」とする．

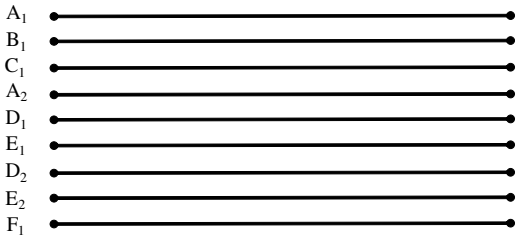

図 3: k-再匿名化を用いた匿名性破綻の解消

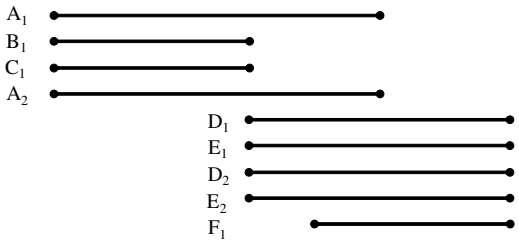

図 4: k^*-再匿名化を用いた匿名性破綻の解消

3.2 k-再匿名化

このようにして再び加工されたレコードは，必ずしも k-匿名性を充足しない．k-匿名性を充足させるためには，同一の準識別子を持つデータ主体が k 以上存在するようにする必要があり，図 1，2 のデータセットは，図 3 のように加工される．このように再び k-匿名性を充足させる操作を k-再匿名化と呼ぶ．

図 3 のレコードは，k-匿名化，複数レコードの突合に対する再符号化 (匿名性破綻の解消)，k-再匿名化の合計 3 回の加工が施されている．そのため，過度に抽象化されており有用性が低下してしまう．

3.3 k^*-匿名性と k^*-再匿名化

そこで，再匿名化に用いる新たな匿名性指標である k^*-匿名性を導入する．

定義 2 *(k^*-匿名性)* あるデータ主体のすべてのレコードの不変準識別子の値は同一であり，かつ各レコードの準識別子の値の組合せを部分集合として持つ他のデータ主体が，*k-1* 以上存在する．

	代替ID	生年	性別	診療月	病名
A_1	A	1970-1976	Any	4月	糖尿病
B_1	B	1970-1974	Any	4月	高血圧症
C_1	C	1970-1974	Any	4月	糖尿病
A_2	A	1970-1976	Any	5月	緑内障
D_1	D	1974-1978	Any	5月	結膜炎
E_1	E	1974-1978	Any	5月	網膜剥離
D_2	D	1974-1978	Any	6月	結膜炎
E_2	E	1974-1978	Any	6月	網膜剥離
F_1	F	1975-1978	女性	6月	緑内障

図 5: k^*-再匿名化データセット

k^*-匿名性を充足させる再匿名化を，k^*-再匿名性と呼ぶ．k^*-再匿名化は，式 2 を用いることで実現できる．k-匿名化したデータセットを k^*-再匿名化した場合，準識別子を知識とした個人特定の可能性は高々$1/k$ である．ただし，元のデータセットに対して，k^*-匿名化のみを施した場合はこの限りではない．k^*-匿名性のみが保証されたデータセットは，全ての準識別子の組合せと，その出現頻度を知識として有している場合に，特定のデータ主体に対応するレコードを k 個未満に特定することが可能である．一方，k-匿名化されたデータセットは，k^*-匿名化のみの場合のような個人特定が生じない．

図 1 のデータセットに対して，k^*-再匿名化を施したデータセットの例を，図 4，図 5 に示す．各直線は，不変準識別子「生年」の値の範囲を表しており，情報損失が少ない．図 4 と図 3 の直性を比較すると，k^*-再匿名化を行った図 4 の方が短く，情報損失が小さいことが分かる．

k^*-再匿名化されたデータセットは同一の準識別子を持つ他のデータ主体を，必ずしも k-1 個持たない．k-匿名化と k^*-再匿名化されたデータセットは，元のいかなるレコードの準識別子の組合せに対しても，それを部分集合とする準識別子を持つデータ主体を k 以上含む．

4 評価実験

本稿の提案手法の有効性を評価するために，評価実験を行った．評価実験では，レセプトデータを匿名化し，情報損失量を評価基準とした．

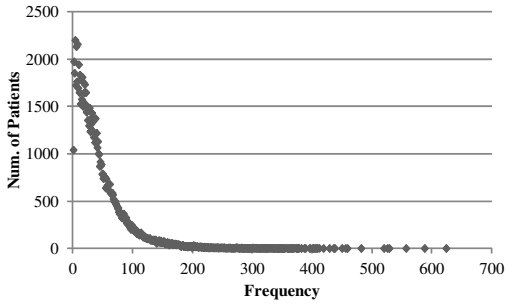

図 6: 評価データセットの頻度分布

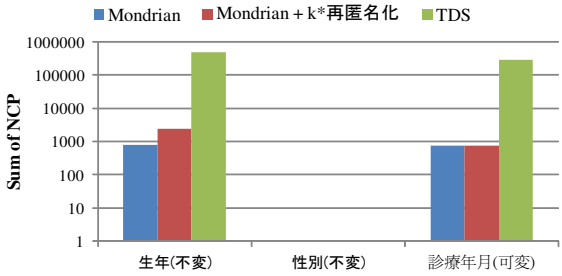

図 7: NCP

4.1 評価環境

評価用のデータセットとして，株式会社日本医療データセンター[1]が提供するレセプトデータを用いた．本稿で用いたレセプトデータは，約 10 万の患者を含む約 400 万のレセプトである．図 6 は，評価データセットにおいて，各患者にレセプトがいくつ存在するかを集計し，その頻度分布を求めた結果である．

評価実験では，k-匿名化 (匿名性破綻あり)，k-匿名化 (匿名性破綻あり)+k^*-再符号化，k-匿名化 (匿名性破綻なし) の 3 方式の比較を行う．k-匿名化 (匿名性破綻あり) として多次元大域的再符号化である Mondrian[4] を利用した．k-匿名化 (匿名性破綻なし) として，TDS (Top Down Specialization)[2] を利用した．Mondrian，TDS ともに，最も汎化された状態から k-匿名性に違反するまで，徐々に汎化された準識別子を詳細化していく手法である．TDS では，ある準識別子の属性について，ある値 qi はすべて同じ qi' に加工される．このため，同一データ主体のすべての不変準識別子が同一の値に加工されるため匿名性破綻が生じないが，情報損失は大きくなる．

4.2 情報損失の評価

本評価では，情報損失の評価のために，属性値の汎化度合いを表す指標である NCP[3] を用いる．NCP 値は 0〜1 の値を取り，0 のときは属性値は全く汎化されていないことを表し，1 のときは最大まで汎化されたことを表す．NCP の定義は以下の通りである．

$$NCP(a) = \frac{|a_{max} - a_{min}|}{|A_i|} \tag{3}$$

a は属性値，a_{max} は a の最大値，a_{min} は a の最小値であり，A_i は a を含む属性の定義域を表す．

本稿では，データセット全体の情報損失を測るために，各レコードの NCP 値の総和を用いて，各手法の情報損失量を比較する．

図 7 は NCP の総和を各手法に対して算出した結果を示している．Mondrian のみを用いた手法は最も NCP の総和が小さく，情報損失が小さい．しかし，この手法では匿名性破綻が生じる．TDS を用いた手法は，匿名性破綻が生じないが，情報損失が非常に大きい．提案手法の Mondrian+k^*-再匿名化は，不変準識別子である「生年」に対しては，Mondrian のみよりも情報損失が大きいが，TDS よりは格段に小さな情報損失である．また，可変準識別子である診療年月に対しては Mondrian のみ，Mondrian+k^*-再匿名化の情報損失は完全に一致している．これは，k^*-再匿名化では，複数のレコードを組み合わせることによる匿名性破綻が生じない可変準識別子に対しては，再匿名化を行わないためである．性別に関しては，どの手法においても汎化されなかった．

以上より，提案手法が複数のレコードを組み合わせることにより発生する匿名性破綻を，小さな情報損失で解消できることが示された．

[1] http://www.jmdc.co.jp/

5　関連研究

データ主体に複数のレコードが存在するデータセットの匿名化はいくつかの研究が存在する．位置情報の時系列データである移動軌跡の匿名化はその一つである．移動軌跡を k-匿名化する (または k-匿名化を拡張した) 手法が提案されている [5][6][7][8]．しかし移動軌跡中の各位置情報は，本稿における可変準識別子であり，複数の組合せから，あるレコードの準識別子の値が詳細化されることはない．また，同一データ主体のレコードが複数のデータセット (リレーション) に存在するときの k-匿名化手法も提案されている [9]．

一方で，レコードのアップデートなどに伴う更新 (re-publication) 時の匿名化が提案されている [10]．しかし，これらの手法は同時に利用できる一データ主体のレコードは一つのみであり，複数のレコードが共通の識別子によって関連付けられることを前提とする本稿とは異なる．

6　おわりに

本稿では，データ主体に複数のレコードが存在するデータセットに対して，複数のレコードを組み合わせることによる匿名性破綻を解消する問題を扱った．匿名性破綻を解消するための再符号化手法を示し，情報損失を抑制できる k^*-匿名性と，それを保証する k^*-再匿名化手法を提案した．k^*-再匿名化は，Mondrian 等の匿名性破綻が生じる一部の k-匿名化手法と併用することで本稿で扱う匿名性破綻を解消できる．評価実験では，Mondrian と k^*-再匿名化を用いた手法が，匿名性破綻が生じない k-匿名化手法よりも小さな情報損失で匿名性破綻の解消を実現できることを示した．なお，k^*-匿名性および k^*-再匿名化に対する数学的証明，および様々な k-匿名化手法との比較評価が今後の課題である．

謝辞

本研究の一部は，経産省の補助事業「平成 23 年度次世代高信頼・省エネ型 IT 基盤技術・実証事業（レセプト情報等の利活用基盤の開発)」プロジェクトの成果である．

参考文献

[1] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp. 555–570, 2002.

[2] B.C.M. Fung, K. Wang and P.S. Yu. Top-down specialization for information and privacy preservation. In Proc. of $ICDE$, 2005.

[3] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi and A. Fu. Utility-based anonymization using local recoding. In Proc. of $SIGKDD$, pp.785–790, 2006.

[4] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity, In Proc. of $ICDE$, 2006.

[5] O. Abul, F. Bonchi, and M. Nanni. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In Proc. of $ICDE$, pp. 376–385. 2008.

[6] M. Terrovitis and N. Mamoulis. Privacy Preservation in the Publication of Trajectories. In Proc. of MDM, pp. 65-72, 2008.

[7] M. E. Nergiz, M. Atozori, Y. Saygin and B. Güç. Towards Trajectory Anonymization: a Generalization-Based Approach. Transactions on Data Privacy, 2, pp. 47–75, 2009.

[8] T. Takahashi and S. Miyakawa. CMOA: Continuous Moving Object Anonymization. In Proc. of $IDEAS$, 2012.

[9] M. E. Nergiz, C. Clifton, A. E. Nergiz. MultiRelational k-anonymity. In Proc. of $ICDE$, pp.1417–1421, 2007.

[10] X. Xiao and Y. Tao. m-invariance: towards privacy preserving re-publication of dynamic datasets. In Proc. of $SIGMOD$, pp.689–700, 2007.