仮名化データの安全性検証アルゴリズムの考察

有住なな[†1]　丹生智也[†1]　南和宏[†1]　丸山宏[†1]

近年、インターネット上では様々な位置情報サービスが提供されており、多数のユーザーの間で位置情報の共有が実現されている．しかし位置情報はユーザーのプライバシーに関する行動に深く関連するため，位置情報の公開は適切に制限される必要がある．そこで我々はこれまでユーザーの識別子を仮名に置き換える仮名化方式を検討し，「ミックスゾーン」と呼ばれる複数ユーザーが出会う場所でランダムに仮名を交換する方式を考案し，各ユーザーが取り得る代替経路の不確定性に基づきプライバシーの概念を定式化した．本論文では、このミックスゾーンを用いたプライバシー保護技術において効率的に安全性を検証するアルゴリズムについて考察する．

1. はじめに

近年、GPS 機能を搭載したスマートフォンが普及し，多くのモバイル･ユーザーが Google Latitude [1]等の位置情報サービスを介して、ID 付きの位置情報を他のユーザーと共有するようになってきた．このような，位置情報は、リアルタイムの交通モニタリング[2]や持続可能な町づくりのための都市計画[3]など、多くの分析に役立つが、その一方で，位置情報はユーザーのプライベートな行動を示唆する可能性が高いため，その機密性の取り扱いには注意が必要であり，位置情報の公開は適切に制限される必要がある．現状では k-匿名化の手法を適用し，地域ごとの人口分布の中でしきい値 k 以上の統計データのみを公開するのが一般的である．しかし k-匿名化の手法では，各ユーザーが時間軸でどのように移動したかを示す軌跡情報が失われ，位置情報の価値を十分生かした分析が可能とはいえない．

そこで我々はこれまでユーザーの識別子を仮名に置き換える仮名化方式を検討し，「ミックスゾーン」と呼ばれる複数ユーザーが出会う場所でランダムに仮名を交換する方式を考案し，各ユーザーが取り得る代替経路の不確定性に基づきプライバシーの概念を定式化した[4]．図１の例ではユーザーp_iとp_jの仮名がランダムに交換され，どちらに移動したかが不確定になる．しかしながら外部知識をもつ攻撃者を想定した場合，その外部知識と一貫性のある代替経路を列挙する問題はそれほど単純でなく，一般には全て（k 人）のユーザーの経路の組み合わせを考慮した k 辺素パスを列挙することが必要になる．k 辺素問題とは，k 組の始点と終点が与えられ，各始点から終点を結ぶ互いに排他的な経路を見つける問題である．

残念ながら，我々が対象とする非循環な有向グラフに対する k 辺素パス問題の計算量は非多項式時間[5]であることが理論的に知られており，我々が過去に行った制約充足問題ソルバー[4]を用いた実証評価でも現実的に解けるデータのユーザー数は１０～２０人程度に留まり、提案手法の適用は小規模の位置情報データに限定されるものであった．

今回我々は、我々が対象とするミックスゾーンのグラフでは、全てのエッジが k 辺素パスに利用されるという特殊な性質に着目し，元グラフを k 辺素パスの探索を効率よく行えるように全てのエッジにラベル付けをする手法を考察した．本考察手法で得られるラベルグラフでは異なる始点・終点間の経路の干渉の可能性を考察する必要がなく，多項式時間である始点・終点間の経路を全て列挙することが出来ることを示した．

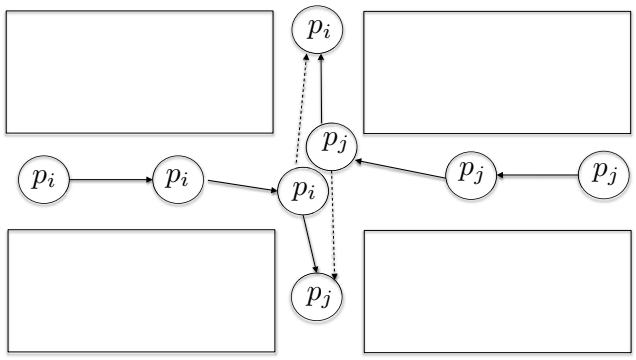

図１．仮名交換の例．ユーザーp_iとp_j交差する地点で仮名がランダムに交換される。

2. プライバシーモデル

図２はある一人のユーザーに注目した時の位置情報を示す．仮名が交換されることで，複数の経路が可能と考えられる．この例では，時間 t においてユーザーが取りえる仮名は多数あり，プライバシーは確保される．しかし，攻撃者はt_0もしくはt_*におけるユーザーの位置を何ならかの外部知識として知りえるとする．例えば時間t_0とt_*の場所がユーザーの自宅であるとすると，そのような情報は名簿から取得可能であり時間t_0とt_*におけるプライバシーは存在しない．それに対して中間点の t においては、ユーザーに関する代替パスが複数存在するのである程度プライバシーは確保できる．しかしミックスゾーンのグラフに対し安全

†1 統計数理研究所

性を定量的に評価する効率的なアルゴリズムが必要となる．我々は，これを有向グラフの k 辺素問題の枠組みで考察することで k 辺素パスを列挙し，すべてのユーザーに対しての安全性を定量化する．

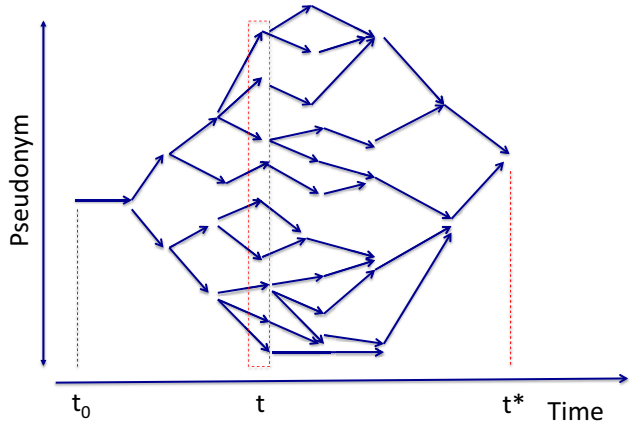

図 2　あるユーザーに対しての仮名パスの例

定義 1．（順次分割グラフ G）．

有向グラフ G = (V, E)において頂点のリスト V を $S_1, S_2, \ldots S_n$ に分割出来，すべての辺 e = vw が S_l と S_{l+1} の間にある場合，G を順次分割グラフという．

図 3 は順次分割グラフの概要を示す．ユーザーは自宅から自宅へと移動すると仮定するため、S_1 と S_n にユーザーの自宅情報が入る。

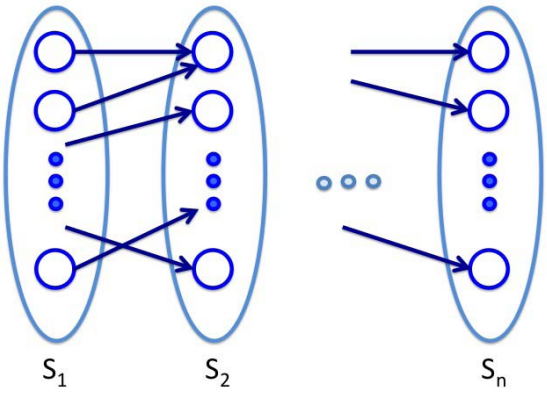

図 3　順次分割グラフの概要

順次分割グラフを用いて，ミックスゾーングラフを定義する．これにより，位置情報の問題をグラフ理論の問題の枠組みに入れることが可能となる．

定義 2．（ミックスゾーングラフ G_k）

順次分割グラフ G = (V, E)　において，k 辺素パス(s_1, t_1), (s_2, t_2), …, (s_k, t_k)　が以下の条件を満たす場合，これをミックスゾーングラフ G_k と呼ぶ．

1. $|S_1| = |S_n| = k$
2. $0 < |S_i| < k$ for all i! {2,3, …, n-1}
3. $p_j^1=(s_i, t_i)=(v^i_1, v^i_n)$ for i! {1,2, …, k}が全ての G の頂点と辺を被覆する辺素なパスが存在する．

3. 辺のラベル付け

全ての辺にユーザーを表すラベルをつけることで，どのユーザーがどの辺を通れるかを示す．ユーザーの仮名が全ての辺について分かっていれば，経路の割り出しは容易になるため，このラベルを多項式時間に割り当てることが必要となる．

定義 3．（ラベル L(e)）

ミックスゾーングラフ G_k = (V, E)に対し，辺のラベルを１から k の整数の集合で表す．

全ての辺に必要でかつ十分なラベルを割り振る必要がある．入力辺のラベルと同じラベルを持った辺の集合をコモンエッジとして定義し，同じラベルを含有する辺の集合を包含コモンエッジとして定義することで、必要以上なラベルが割り当てられていることを発見しやすくする．この際，頂点 v に入ってくる近傍を $N^-(v)$，出て行く近傍を $N^+(v)$として表記する．

定義 4．(コモンエッジ CE(v,e))

ミックスゾーングラフ G_k = (V, E)に対し，コモンエッジを頂点 v と辺 e に対し、以下のように定義する．

$$CE^-(e,v) = \{wv | w \in N^-(v) \wedge L(wv) = L(e)\}$$

$$CE^+(e,v) = \{vw | w \in N^+(v) \wedge L(vw) = L(e)\}$$

また、包含コモンエッジを頂点 v と辺 e に対し、以下のように定義する．

$$CE_i^-(e,v) = \{wv | w \in N^-(v) \wedge (\exists l \in L(wv) : l \in L(e))\}$$

$$CE_i^+(e,v) = \{vw | w \in N^+(v) \wedge (\exists l \in L(vw) : l \in L(e))\}$$

次に，全ての辺にラベルをつけ，そのグラフを定義する．ユーザーが頂点を移動するため、辺素パス(s_i, t_i)に対してインデックス i をユーザーの仮名として用いる。

定義 5．(ラベルグラフ $L(G_k)$)

ミックスゾーングラフ G_k = (V, E)に対し，ラベルグラフ $L(G_k)$を以下のように定義する．

1. S_1 から出てくる全ての辺に対して、ラベルはインデックスと同じである。
   $\forall e = v_i^1 w \ for \ v_i^1 \in S_1 : L(e) = i.$
2. S_n から入る全ての辺に対して、ラベルはインデック

スと同じである。

$$\forall e = vw_i^n \ for\ w_i^n \in S_n : L(e) = i.$$

3. 1.2.に含まれない全ての辺に対して、ラベルは両辺のラベルによる. 辺 e に入る辺を In(e), 出る辺を Out(e) とすると,

$$\forall e = vw\ where\ v \in S_j\ \forall j \in \{2,3,\ldots,n-2\} :$$
$$L(e) = (\bigcup_{e' \in In(e)} L(e')) \bigcap (\bigcup_{e' \in Out(e)} L(e')).$$

4. 頂点に対し，ラベルの数を見ることで，必要のないラベルを消すことが出来る.

$$\forall e = vw\ where\ v \in S_j\ \forall j \in \{1,2,\ldots,n-1\} :$$

において,以下の４つの条件を一つでも満たす場合頂点に入る同じラベルの辺の数は同じになる.

(a) $CE_i^-(e,w) = CE^-(e,w) \wedge CE_i^+(e,w) = CE^-(e,w)$

(b) $CE_i^+(e,v) = CE^+(e,v) \wedge CE_i^-(e,v) = CE^+(e,v)$

(c) $|CE_i^-(e,w)| = |CE_i^+(e,w)|$
$\wedge \exists x$ such that $\forall e' \in CE_i^+(e,w)\, e' = wx$
$\wedge \exists A \forall e' \in CE_i^+(e,w)$
where $L(e) \in L(e')$ such that $L(e') = A$

(d) $|CE_i^-(e,v)| = |CE_i^+(e,v)|$
$\wedge \exists x$ such that $\forall e' \in CE_i^-(e,v)\, e' = wx$
$\wedge \exists A \forall e' \in CE_i^-(e,w)$
where $L(e) \in L(e')$ such that $L(e') = A$

$$\Longrightarrow |CE^-(e,w)| = |CE^+(e,w)|,$$

辺のすべてにラベルがついているラベルグラフを用いて，ガイデッドパスを見つけることが出来る．同じラベルの経路が複数出来ることで、図２に示したような仮名パスのループが出来る。

このようなラベル付けが，あるユーザーのすべての可能経路を示す.

定理 1. ミックスゾーングラフ $G_k = (V, E)$が与えられているとすると，ラベルグラフ $L(G_k)$にあるガイデッドパスを取り除いたグラフは必ずミックスゾーングラフになる.

証明 数学的帰納法を用いて証明する．まず，定理は全ての k＝n について成り立つと仮定して,ミックスゾーングラフ G_{k+1}についても定理が成り立つことを示す.

条件１：k=1 の場合は，経路は一つだけになり，これを取り除くとグラフはなくなり，条件を満たす.

条件２：k＝n+1 の場合を考える．もし，定理が正しいならば, $L(G_{n+1})$にあるどのガイデッドパス p^l_jを取り除いても，残りがミックスゾーングラフになる．これは，あるガイデッドパス q が $L(G_{n+1})$にあり，q を除いたグラフもミックスゾーングラフであり，かつ p^l_jが存在していることと同義である．そこで，そのような q が存在していないことを仮定して，矛盾があることを示す.

G_{n+1} がミックスゾーングラフであるため，辺素パスが k あり，$p^1{}_j$が必ず存在する．そこで，まず $p^1{}_j$を G_{n+1}から取り除く．$p^1{}_j$は辺素パスのため，これを抜いた G_nもミックスゾーングラフになる．この G_n にあるガイデッドパス q を取り除いても,数学的帰納法から残りは,ミックスゾーングラフ G_{n-1} となる．$p^1{}_j$ のパスを戻すことで，図４に示すように,新しいミックスゾーングラフを作ることが出来る．この時，p^l_jが存在していれば，q も存在することになり，仮定に反する．そのため，q のラベル i があるパス $p^m{}_i$は全て必ず p^l_jとどこかで同じ辺を共有していることになる.

$$G_{k+1} \xrightarrow{\text{Remove } p_j^1} G_k \xrightarrow{\text{Remove } q} G_{k-1} \xrightarrow{\text{Add } p_j^1} G'_k$$

図 4　ミックスゾーングラフの変換

このような全てのガイデッドパスと p^l_j が共有の辺を持った G_nでも、G_{n+1}で q と $p^1{}_j$の経路と一部置き換えることが出来れば、その経路 q'は G_{n+1}に存在する．かつ，q'を取り除いても $p^1{}_j$が q と交換した経路も残っているため，ミックスゾーングラフとなっている．このような q'があると仮定に反するので，このような q'がないケースを考える必要がある.

そこで，全てのガイデッドパスと p^l_j が共有の辺を持った G_nは存在しないことを示す．G_{n+1}において，すべてのガイデッドパスと p^l_j が共有の辺を持つ場合は，強い条件となる．そこで，この条件が成り立たないことを背理法で示す．もし，この条件が成り立つならば，ある辺のラベルに i と j が両方あり，かつ，この辺に i がなければ条件がなり立つ．そこで，この辺を e_1と呼ぶ．この辺に複数のラベルがあることより，i はどこかから分岐していると言える．そうでなければ，ラベルグラフの条件 4 に違反する．そこで，e_1がある区間にある i が全て分岐する点があり，そこからの全ての i のパスに p^l_j と共有する辺がある.そのような辺の一つを e_2とする。e_1 と e_2は両方とも p^l_jに含まれるため，e_1 と e_2を繋ぐ経路 X には j のラベルが含まれる．ただし，e_1 と e_2は別の経路であるため，i のラベルは含まれない．ラベルグラフの条件 1，2，3 だけでは，e_1 ，e_2，X 共に i と j のラベルを含みラベルグラフの条件４で i のみを X から取り除くことが可能でなければならない.しかし，e_1 ，e_2には i と j のラベルが必要なため，ラベルグラフの条件４だけでは，X から i がなくならない．そのため，背理法の条件を見たし，G_{n+1} は存在しない.. このことから，G_nが存在せず，q の存在が示された.

これにより，G_{n+1}から，どのような p^l_jを取り除いても，必ずミックスゾーンになると言える.

4. 安全性検証アルゴリズム

ラベルグラフを作るアルゴリズムが存在すれば，定理１より，すべてのユーザーについての可能な経路が導かれ,

安全性の検証が出来る。

以下のパラメーターを用い，安全性検証にかかる計算量が多項式時間であることを求める．

1. ユーザー数：k
2. 時間ステップ：T

ラベルグラフが与えられている場合，可能な経路の検索は，深さ探索アルゴリズムで求まるので，O()

そのため，アルゴリズム 1 が多項式時間で与えられるならば，安全性検証が多項式時間で行える．

定理 2． ミックスゾーングラフ $G_k = (V, E)$が与えられているとすると，$O(k^3T^2)$のアルゴリズムが存在する．

証明 ベースケースでは，一つの区間内で k のビットオペレーションを最大 k 回行うので全部で $O(Tk^2)$かかる．

再起ケースでは，一つの頂点に l の辺が入るならば，コモンエッジを見つけるのに O(kl)かかる．そこで，一つの区間内では，$O(kl^2)$かかり，全体では

$\mathcal{O}(\sum_{i=1}^{|S_i|} kl_i^2) < \mathcal{O}(k(\sum k)^2) = \mathcal{O}(k^3)$

かかる．

もし，一つの頂点でラベルのうち n が変わるとすると，最大で k-n の辺が変わり，区間内では，

$\mathcal{O}((k-n)nk) < \mathcal{O}(k^3)$

かかる．そのため，全体では $O(k^3T)$かかる．

あるユーザーで見れば，一回の再起で必ず一つの区間は安全経路が発見されるため，全体で O(T)で収束する．そのため，全体では $O(k^3T^2)$かかる．

このような再起法を用いたアルゴリズムを作る必要がある．まずベースケースを与える．

Algorithm 1 base case

```
COMMENT: Set condition 1, 2, and 3
function BaseCase(G = (V, E))
    COMMENT: set everything to empty
    for ∀e ∈ E do
        L(e) ← ∅
    COMMENT: from the start
    for each e = vw such that v ∈ S_1 and v ∈ S_2 do
        L(e) ← index of v
    for i ← 2 to n − 2 do
        for each e = vw such that v ∈ S_i and v ∈ S_{i+1} do
            L(e) ← ∪_{e'∈In(e)} L(e')
    COMMENT: from the start
    for each e = vw such that v ∈ S_{n−1} and v ∈ S_n do
        L(e) ← index of endpoint in S_n
    for i ← n − 2 to 2 do
        for each e = vw such that v ∈ S_i and v ∈ S_{i+1} do
            L(e) ← ∪_{e'∈Out(e)} L(e')
```

これで，ラベルグラフの条件 1, 2 , 3 は満たされるので，再起法を用いて条件 4 を満たすように，必要でないラベルを消し，そのアップデートの後も条件 3 を満たすように変えて行く．まず，ラベルグラフの条件 4 を満たしているか判断する必要がある．これには，時間軸沿いに 2 種あるので，下記のアルゴリズムを使う．

Algorithm 2 condition1

```
COMMENT: Check if condition 4 holds or not
function Condition1(CE_i^-, CE_i^+, CE^-, CE^+)
    if |CE_i^-| == |CE^-| ∧ |CE_i^+| == |CE^+| ∧ |CE^-| < |CE^+| then
        return true
```

Algorithm 3 condition2

```
COMMENT: Check if condition 4 holds or not
function Condition2(L, CE_i^-, CE_i^+, CE^-, CE^+)
    if |CE_i^-| == |CE_i^+| then
        if all edges in CE_i^+ has the same endpoints then
            if all edges in CE_i^+ which includes L has the same labels then
                return true
```

これを使って，全体のアルゴリズムを作る．

定理 3． ミックスゾーングラフ $G_k = (V, E)$が与えられているとすると，以下のアルゴリズムがラベルグラフを作る．

Algorithm 4 full algorithm

```
function Algorithm(G = (V, E))
    BaseCase(G)
    while any label changed do
        for i ← 2 to n − 1 do
            COMMENT: update according to condition 3
            for each e in between S_i and S_{i+1} do
                if In(e) was updated then
                    L(e) ← L(e) ∩ ∪_{e_in∈In(e)} L(e_in)
            for v in S_i do
                if any adjacent edges were updated then
                    COMMENT: Find common edges
                    CE^+ ← CE^+(uv, v)
                    CE_i^+ ← CE_i^+(uv, v)
                    CE^- ← CE^-(uv, v)
                    CE_i^- ← CE_i^-(uv, v)
                    if Condition1(CE_i^-, CE_i^+, CE^-, CE^+) then
                        all labels of edges in CE^+ to be L(uv)
                    if Condition2(L(uv), CE_i^-, CE_i^+, CE^-, CE^+) then
                        all labels of edges in CE^+ to be L(uv)
        for i ← n − 1 to 2 do
            Same as above but from the other direction
```

証明 ベースとして，ラベルグラフの条件 1 と 2 が当てはめられている．これは，再起的に変わることはないため，条件 1 と 2 を満たす．このアルゴリズムでは，ステップごとにラベルの数は減らないため再起は必ず収束する．かつ，一つ一つの頂点に対してラベルグラフの条件 3 と 4 を当てはめているので，収束後は，条件 3 と 4 を満たす．

5. 結び

本論文では，位置情報の移動パターンによる推論を利用したミックスゾーングラフを有向グラフの k 辺素パスの問題で定式化することで，安全性の考慮ができることを示した．必ず同じ長さの k 辺素パスが存在することを利用し，位置情報の安全性を保証する手法を考察した．

参考文献

1) Google latitude. http://www.google.com/latitude.

†1 統計数理研究所

2) Google maps. http://www.maps.google.com/.
3) Seike, T., Mimaki, H., Hara, Y., Odawara, R., Nagata, T., Terada, M.: Research on the applicability of “mobile spatial statistics” for enhanced urban planning. Journal of the City Planning Institute of Japan 46(3), 451–456 (2011)
4) Tanjo, Tomoya, et al. "On Safety of Pseudonym-Based Location Data in the Context of Constraint Satisfation Problems." *Information and Communication Technology*. Springer Berlin Heidelberg, 2014. 511-520.
5) Fortune, Steven, John Hopcroft, and James Wyllie. "The directed subgraph homeomorphism problem." *Theoretical Computer Science* 10.2 (1980): 111-121.
6) Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In Proceedings of Mobisys 2003: The First International Conference on Mobile Systems, Applications, and Services, San Francisco, CA, May 2003. USENIX Associations.
7) Urs Hengartner and Peter Steenkiste. Access control to people location information. CM Transactions on Information and System Security (TISSEC), 8(4):424–456, 2005.
8) Ginger Myles, Adrian Friday, and Nigel Davies. Preserving privacy in environments with location-based applications. IEEE Pervasive Computing, 2(1):56–64, January-March 2003.