Vol.2013-EIP-62 No.12
2013/11/21

滞在場所の k-匿名化と濡れ衣

中川裕志[†1]　　角野為耶[†1]

ビッグデータ利用において個人特定をできることがプライバシー保護の観点から問題であることが指摘され，プライバシー保護データマイニングの分野で匿名化手法の研究が進んできた．一方，個人が自分自身と直接関係しない情報で否定的な事情を疑われること，すなわち濡れ衣の被疑はこれまでプライバシー保護データマイニング分野では見過ごされてきた問題である．本研究では，このような問題意識から特定の事象と無関係な人にその事象に関する疑いがかかる濡れ衣現象を誘発する構造を明らかにし，対策を検討する．

k-Anonymization of Locations and False Light

HIROSHI NAKAGAWA[†1]　NASUKA SUMINO[†1]

Identifying person in using the Big Data is problematic in privacy protection, resulting in the advancement of anonymization methods for privacy preserving data mining. On the other hand, a person is misleadingly suspected as a bad guy due to the information which actually has nothing to do with him/her. This is so called false light which we have not paid enough attention so far. In this report, we identify the mechanism which generates false light and also propose one of countermeasure.

1. 濡れ衣の被疑

インターネット上のデータや諸機関が収集した大規模データ（ビッグデータと呼ばれる）については，その有効な利用を巡る議論が活発になされている．しかし，個人に係わる情報である場合，ビッグデータ利用において個人特定をできることがプライバシー保護の観点から問題であることが指摘され，プライバシー保護データマイニングの分野で匿名化手法の研究が進んだ．一方，個人が自身に直接関係しない情報で否定的な事情を疑われること，すなわち濡れ衣の被疑はこれまで見過ごされてきた問題であろう．以下では，このような問題意識から匿名化と濡れ衣の関係について考察する．

1.1 濡れ衣の被疑の実例：人名検索

個人Aが濡れ衣を着せられるとはA以外の人の犯罪行為をAの犯罪とみなされてしまうことであり，法律的には冤罪ということである．だが，濡れ衣を着せられなくても，濡れ衣の被疑は潜在的にもっと多いであろう．

ビッグデータ時代における濡れ衣の被疑の実例としては，以下の Google サジェストによる不利益に関する判例がある．原告は Google に自分の名前を検索キーワードとして入力すると，サジェスト機能によって犯罪を連想させる語句が表示され，就職活動に支障をきたしたとして名誉毀損で Google を訴えた．同種の訴訟が 2 件あった．

第 1 のケースは 4 月 15 日に東京地裁で表示差し止めを命じた．すなわち「サジェスト機能により，名誉毀損に当たる投稿を見る人がおり，放置すれば権利侵害が拡大する」と判断した．

第 2 のケースでは 5 月 30 日に原告の訴えを退けた．すなわち，「検索結果自体から，サイトの内容が名誉やプライバシーを侵害し社会通念上容認できないものか一見して明らかとは言えない」と指摘している．

両方のケースとも控訴したらしく，東京高裁で係争中と言われている．これらのケースは Google のサジェスト機能により，犯罪と関係があるのはないかという疑いをうけたもので，上記の濡れ衣の被疑の例といえる．ちなみに，この問題は人名検索の精度が悪いことに起因する部分があるので，人名検索の場合は人名検索の結果を実世界の異なる人物毎にクラスタ化してサジェスト機能に適用すれば防げるはずである．つまり，同姓同名の参照曖昧性を完全に解消できれば問題は技術的に解決する．では，人名検索の曖昧性解消の技術はどの程度のレベルかを次にみてみよう．

人名検索結果の Web ページを現実世界での同一人物毎にクラスタ化する Web 名寄せタスクは WePS と呼ばれる競争型タスクが WePS1，WePS2 と 2 回行われた．WePS2 は 2009 年に行われ，最高性能は香港科学技術大学の F 値 0.81 であった．我々は第 2 位で F 値は 0.80 であった[4]．つまり，現在の技術レベルでは，上記のサジェスト機能が濡れ衣の被疑を誘発しないようにすることはできない．

この例以外の濡れ衣被疑で当事者に不利なものとしては以下のような例を容易に想像できる．

- 消費者金融に行ったという濡れ衣．
- ドラッグを買ったのではないかという濡れ衣．

上記 2 例は、就職，婚活で不利になるかもしれない．

- 感染症で病院に行ったのではないかという濡れ衣．

当人にせいではないが，職場や友人関係で不利なこと

†1 東京大学
The University of Tokyo

があるかもしれない．

もちろん，間違いにしても当事者にとってうれしい例もあり，むしろ，良いことで宣伝したい場合もあるかもしれない．例えば，

- タレントの握手会に行った．
- チケットが入手しにくい試合を見た．

などである．

1.2 k-匿名化を巡る問題

インターネットやビッグデータにおいて、仮に個人名を消した匿名化を行っても、名前以外の情報（郵便番号、年齢、居住地、性別など。これらは疑似 ID と呼ばれる。）から個人が特定される問題は 2000 年以降さかんに指摘された．その対策として Sweeny によって提案された *k*-匿名化[3]は，事象 S の当事者が一意に識別されないように，事象 S と関係する当事者を含むデータベースに改変を加える方法である．表 1 のようなデータベースを例にして説明する．

このデータベースでは個人識別フィールドの名前は匿名化されている．匿名化された名前(A,B,C,D)を仮名と呼ぶ．誕生日，性別，郵便番号という疑似 ID を組み合わせると仮名と一意の対応が付き，病名が分かる．病名のような個人にとって重要な情報が記載されたフィールドは機微な属性(sensitive feature)が記載されたフィールドだが，ここでは問題事象フィールドと呼ぶことにする．

表 1 データベース

Table 1 Data Base.

仮名	誕生日	性別	郵便番号	病名
A	21/1/79	男	53715	インフルエンザ
B	10/1/79	女	55410	風邪
C	21/2/83	男	02274	口内炎
D	19/4/82	男	02237	インフルエンザ

もし，別のデータベースで疑似 ID と個人名の対応が分かれば，問題事象フィールドに記載された機微な個人情報である病名が特定されうる．例えば，米国では選挙人名簿が公開されているため，個人特定は 90%に近い確率で可能である．そこで，k -匿名化では，疑似 ID の内容のデータの精度を落とすことによって，疑似 ID のデータを組み合わせても，同じ疑似 ID の組み合わせを持つ人が少なくとも k 人以上存在するようにして，個人を特定できないようにする．

表 2 2-匿名化

Table 2 2-Anonymization

仮名	誕生日	性別	郵便番号	病名
A	*/1/79	*	5****	インフルエンザ
B	*/1/79	*	5****	風邪
C	*/*/83	男	022**	口内炎
D	*/*/82	男	022**	インフルエンザ

例えば，表 2 のように疑似 ID フィールドのデータの一部を*で置き換えると 2 人以下には絞り込めないので 2-匿名性が実現できる．なお，このような改変を行ったとき，データマイニングにおいて、どの程度の有用な情報が取り出せるかは，応用分野依存である．

1.3 位置情報の匿名化と危険性

ビッグデータの中でも価値が高いとされているのが個人の行動履歴である．地理的行動履歴は，公共機関の設計や販売戦略に役立つと言われている．しかし，個人情報が流出することの危険性の指摘は既にあり，[5]でも，Web 検索や SNS などを使うと SNS 上の交友関係から個人が特定される例が報告されている．この報告とは異なり，学問的裏付けがない場合でも，個人の行動履歴情報については，その利用に関して漠然とした危惧をいだく人が多数であろう．例えば，2013 年 7 月に JR 東日本が Suica の行動履歴データを匿名化した上で日立製作所を経由して販売するという報道がされたとたんに，反対意見や多くの懸念が SNS などで噴出し，販売計画は中止されている．一般大衆の心理は正確には分からないが，匿名化が万全ではないこと，Suica 購入時にこのようなデータ利用法が知らされていなかったことから，不信感が露わになったものと推測される．

したがって，個人の行動履歴などをビッグデータとして活用するための社会的コンセンサスを得るためには，ビッグデータを使った場合の危険の程度，対策などを技術的に把握し，周知することが必須となる．その上で，ビッグデータ利用の社会的コンセンサスに至るのが理想であろう．

位置情報の k -匿名化は位置情報の最小単位を点ではなく領域と考え，その内部に k 人滞在するような広さに領域を拡大することによって実現する．当然，k が大きくなると領域が拡大し，情報の位置精度が劣化する．技術的には領域の形状を工夫することによって k -匿名化を実現して，可能な限り情報の位置精度を確保することが目的になる．

1.4 *l*-多様性と *t*-近接性を巡る問題

疑似IDフィールド群の値が同一である k 人が全員同一の問題事象フィールドの値を持っていると k -匿名化の意味がない．そこで，k 人中，問題事象フィールドの値は少なくとも l 種類を確保するのが *l*-多様性[2]である．しかし，*l*-多様性では次のような悪影響が指摘されている．

簡単のために 3-匿名性で 2-多様性，すなわち k=3, l=2 とする．問題事象フィールドが病名であったとする．2-多様性を確保するために疑似 ID が同一の 3 名中、1 名が深刻かつ潜伏期間の長い感染症Ｚ，2 名が風邪となったとする。すると、以下の 2 つの問題が起こる．

(a) 外部のデータベースとの突き合わせで 2 名が風邪であると分かると，残り 1 名がＺであることが暴露されてしまう．

(b) 従来は問題視されてこなかったが，3 名の中にZの患者が入っていることが分かると, 風邪の 2 名は 1/3 以上の確率でZであることになるので，Zの感染者という濡れ衣の嫌疑を受けてしまう．

このように，*l*-多様性は副作用があるため，それを緩和する方法として *t*-近接性[1]が提案された．*t*-近接性は疑似 ID が同一の *k* 人において，問題事象フィールドの値の分布がデータベース全体での分布にできるだけ近くなるようにしようというものである．*k* が非常に大きければ，上記の *l*-多様性の問題(a)(b)は緩和される．しかし，*k* =5 程度では問題は緩和されない．*k* =5 でZが 1 名入っていたとしよう．他の 4 名が全てZでないことが分かる可能性が *k* =3 の場合よりは緩和されるので、(a)の問題は大幅に緩和される．しかし，(b)は依然として問題である．例えば，Zの患者は母集団での感染者確率が 0.01%だったとすれば, *k* =1000 程度にしないと *t*-近接性の効果はない．1000 人中１名しかZがいなければ, 濡れ衣嫌疑は非常に薄いものになる. しかし，*k* =5 であると，Zではない 4 名もZの患者である確率は 20%であるので，疑いの目に晒される．結局，対象の人物がZの患者であることが，自分たちにどの程度の影響があるのかで濡れ衣嫌疑の確率が変わってくる．2 章では，この濡れ衣嫌疑の確率について議論する．

2. 濡れ衣被疑の構造

既に述べた例のように病院に行ったことが知られることによって感染症の患者であることを疑われる濡れ衣の被疑は，従来，ビッグデータとの関連は注目されてこなかったが，その発生のメカニズムを明らかにしておくことは意味がある．

2.1 濡れ衣の形式的記法

濡れ衣とその被疑についての形式的な記法を説明する．ビッグデータやインターネット上の情報環境では，与えられた情報環境 *C* において議論を進める必要がある. そこで，ここでは情報環境 *C* が与えられたとき, 主体 *A* が個人 *X* に対して事象 *S* の当事者である疑いを持つか否かを表わす論理式を

$$\mathrm{f}(A,\mathrm{t}(X,S),C) \tag{1}$$

で表わす．情報環境 C とはビッグデータに含まれる情報，あるいは外部から与えられた情報を意味し，具体例は後に述べる．

与えられた情報環境 *C* において，A が *X* の *S* を疑う度合い、すなわち主観確率を次式で表わす．

$$p_{sub}(\mathrm{f}(A,\mathrm{t}(X,S),C)) \tag{2}$$

ただし，特定の主体 *A* を考慮すると一般的ではないので，(2)を *A* で周辺化した(3)を今後は使う．

$$E_A[p_{sub}(\mathrm{f}(A,\mathrm{t}(X,S),C))] = p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C)) \tag{3}$$

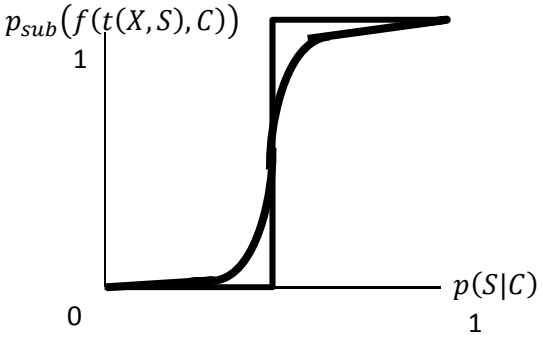

図 1 インフルエンザの被疑の主観確率
Figure 1 False light subjective probability of influenza

個人 X に対して事象 S の当事者であるか否かを論理式 $\mathrm{t}(X,S)$で表わす．すなわち

if X が S の当事者である　then $\mathrm{t}(X,S)$
otherwise $\neg\mathrm{t}(\mathrm{X},\mathrm{S})$ (4)

一方，

if C において A が X に S の疑いを持つ
then $\mathrm{f}(A,\mathrm{t}(X,S),C)$
otherwise $\neg\mathrm{f}(A,\mathrm{t}(X,S),C)$ (5)

したがって，*X* が *A* に *S* の濡れ衣の疑いを掛けられたことは次式で表わされる．

$$\neg\mathrm{t}(X,S) \text{ and } \mathrm{f}(A,\mathrm{t}(X,S),C) \tag{6}$$

したがって，濡れ衣被疑の主観確率は

$$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))p(\neg\mathrm{t}(X,S)) \tag{7}$$

となる．当然ながら，X が S の当事者であって，かつそのことを疑われる主観確率は次式である．

$$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))p(\mathrm{t}(X,S)) \tag{8}$$

濡れ衣の被疑を実際受けてしまえば，その被害額や救済法ということが問題になるが，それは法律的ないし社会制度的な問題なので，ここでは扱わない．ビッグデータやインターネット上の情報環境で問題になるのは，式(3)で表わされた主観確率である．しかし，これまでは式(3)の濡れ衣の被疑の主観確率を小さくしようという問題意識は示されてこなかった．

2.2 濡れ衣の被疑を誘発する認知の構造

まず，情報環境 *C* が与えられたとき，事象 *S* が発生している確率を $p(S|C)$ とする．例えば，インフルエンザの患者の多い病院において，インフルエンザ患者数が通院した 100 人中 3 名ということが *C* として与えられていたとしよう．すると，通院していた患者がインフルエンザであるという事象 S の確率 $p(S|C) = 0.03$ である．

さて，インフルエンザにかかってしまうと，痛みや熱で苦しんだり，勤務先や学校を 1 週間くらい休んだりしなければならず，損害$L(S)$が発生する．インフルエンザ患者と接触したときうつされてしまう確率をp_{trans}とする．この

とき病院に行った人と接触したことによって被る損害額の期待値は

$$\mathrm{E}[L(S)] = p(S|C) \cdot p_{trans} \cdot L(S) \qquad (9)$$

である．一方，少し熱が出たというような状況で病院に行き半日，仕事をしないなどということで発生するインフルエンザ対策コストをCostとする．すると，経済的効果を元に行動するなら式(9)の損害額の期待値 $\mathrm{E}[L(S)]$が Cost より大きければ病院にいくなどの対策を実行するであろう．したがって，病院に行った人にインフルエンザであると疑う主観確率$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))$は，自分に降りかかる損害額の期待値と対策のコストの小さい方を選ぶという最適化をするなら以下の式，あるいは$p(S|C)$を横軸にとった場合は図 1 の階段状の折れ線で示したものになる．

$$\text{if } \mathrm{E}[L(S)] < \mathrm{Cost} \text{ then } p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))=0$$
$$\text{otherwsie } p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C)) = 1 \qquad (10)$$

あるいは，**1**(conditon)を condition が成立したときだけ 1 で，それ以外では 0 となる identity 関数とすれば，次式と表せる．

$$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C)) = \mathbf{1}\left(p(S|C) > \frac{\mathrm{Cost}}{p_{trans} \cdot \mathrm{L}(S)}\right) \qquad (11)$$

もっとも，現実世界では，インフルエンザ罹患確率が小さくても，疑いを持ち対策を実行する人もいれば，インフルエンザ罹患確率が相当大きくても楽観視して対策を行わない人もいる．よって，実際は図 1 においてＳ字に近い曲線で描かれたシグモイド関数風な形になるであろう．

図 1 の$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))$ は病院に行った人がインフルエンザ患者である疑われる主観確率であり，C を「病院に行った」という文脈情報とする．

3. 匿名化と濡れ衣の被疑

3.1 濡れ衣被疑の例

匿名化の有力名手法である k-匿名化が誘発する濡れ衣の被疑を説明するために表 3 のデータベースについて考えてみる．

表 3 所在場所のデータベース例

Table 3 Example data base of location

名前	年	性	住所	N 月 M 日 P 時の所
一郎	35	男	文京区本郷 XX	K 消費者金融店舗
次郎	30	男	文京区湯島 YY	T 大学
三子	33	男	文京区弥生 ZZ	T 大学
四郎	39	男	文京区千駄木	Y 病院

最左列は人名だが，これは匿名化されなければならない．2，3，4 列は，疑似識別フィールドと呼ばれ，1~4 列が総合されると，就活や婚活中に人にとっては，最右列の所在地に消費者金融が記載されていることは芳しくない．そこで，名前を A，B，C，D と仮名化し，疑似識別フィールドの情報を粗いものに変更して表 4 のように改変する．

表 4 4-匿名化したデータベース

Table 4 4-anonymized data base

仮名	年齢	性別	住所	N 月 M 日 P 時の所在
A	30 代	男	文京区	K 消費者金融店舗
B	30 代	男	文京区	T 大学
C	30 代	男	文京区	T 大学
D	30 代	男	文京区	Y 病院

こうすると疑似識別フィールドは 4 人とも同じになるので，4-匿名化が実現でき，消費者金融に行った人を特定できない．事象 S を消費者金融に行ったこととし，情報環境 C を 4 人同一の疑似識別フィールド情報を持つ人がいるとすれば，$p(S|C) = 1/4$であるので，式(3)に示される主観確率で消費者金融に行ったという濡れ衣の被疑を他の 3 人が受けることになる．k-匿名化の k を大きくすると，$p(S|C)$が小さくなるので、事象 S の当事者は特定されにくくなる．同時に濡れ衣を疑われる主観確率は減少するが、濡れ衣の被疑者数も増大する．

3.2 濡れ衣の被疑の減少法

図 1 に示された式(11)の$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))$の構造をみると，$p(\mathrm{S|C}) < \frac{\mathrm{Cost}}{p_{trans} \cdot \mathrm{L}(S)}$ の部分で$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C)) \ll 1$ となる．$p_{sub}(\mathrm{f}(\mathrm{t}(X,S),C))$は$p(\mathrm{S|C})$が十分小さいとき非線形に減少する．すなわち，図 1 に示されたように，$p(\mathrm{S|C})$に対して階段状あるいはＳ字型曲線で増加するので，$p(\mathrm{S|C}) < \frac{\mathrm{Cost}}{p_{trans} \cdot \mathrm{L}(S)}$ となるような大きな k を用いる，換言すれば小さな$p(\mathrm{S|C})$になるような k-匿名化をすれば濡れ衣被疑の主観確率は 0 に近づくほどに低減できる．k 人を含む領域中に存在する S の当事者 s が小さくなるように領域の範囲を定めれば$p(\mathrm{S|C}) = s/k$ が小さくなり，上記の目的を満足する．仮に病院や店にように狭い領域に事象 S の当事者が s_0 人いるなら，後に示す図 2 のようにこの領域を分割して s/k を小さくする領域の構造を求めることが解決策となる．s は「領域の構造」によって決まるので，以後，領域を意識する必要があるときは s(領域の構造)と書く．

3.3 濡れ衣の被疑の減少する領域の拡大法

個人が存在した場所の情報からの濡れ衣被疑の主観確率を低減する手法で領域の形状を変化させる場合のヒューリスティックについて述べる．

まず，事象 Sに関する領域が広がりをもった地区である場合について考える．たとえば，ドラッグ取引の横行する地区は特定の街路の南側と知られているとしよう．この場合は k -匿名化を行うために領域を拡大するときの制約条件として，

・領域はこの街路を越えて拡大してはいけない

とすれば，よい．こうすれば，街路の北側にいる人々がドラッグ取引の濡れ衣の被疑を受けることはない．

次に病院のような特定に施設に滞在した場合について考えよう．この場合は，そもそも施設に滞在した人数が多い．病院そのものに行ったことを隠したい情報とすれば，病院を含む大きな領域を作り k-匿名化するためには，極端に大きな領域を作らなくてはならず，データの利用価値が著しく低下しかねない．そこで，図 2 に示すように，病院を m 個（この図では 4 個）に分割し，分割された病院の一部分に滞在した人数を減らした上で領域の拡大を行う方法が考えられる．

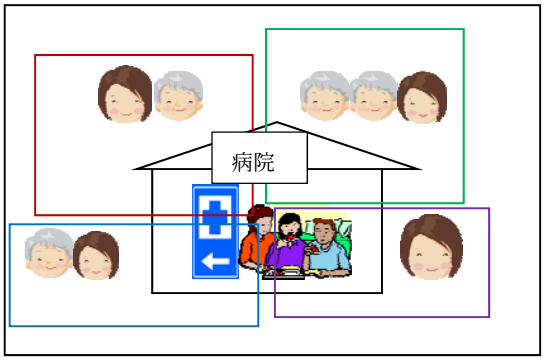

図 2 病院を 4 分割して拡大した領域
Figure 2　Enlarged areas by dividing the hospital into 4 areas.

図中の四角形の中には k 人滞在し，病院の滞在者数 s 人のうち，$s/4$ 人だけが一つの領域に入っている．この手法を施設分割あるいはポイント分割ヒューリスティックと呼ぶことにする．

ポイントとなる病院、貸し金融店舗、裁判所、警察署、風俗店などを上記のように分割すると、k -匿名化は、まずこれらのポイントを含む k 人以上内部に滞在する k-匿名化の領域を作りことになる。こうして作られた k-匿名化の領域でカバーされていな場所を適当に分割して k-匿名化すればよいので、一種の分割統治型のアルゴリズムになる。ただし、異なるポイントを含む 2 個以上の k-匿名化領域が重複した場合の処理は工夫が必要であり、今後の課題である。

3.4 l-多様性と t-近接性

k-匿名化された領域の k 人中，問題事象フィールドの値は少なくとも l 種類を確保するのが l-多様性であった．ただし，$k > l$ する．以下の両極端な場合を考えてみる．問題事象 S を持つ人は最大 k -l-1 人である．すると，

$$p(S|C) = (k - l - 1)/k \qquad (12)$$

逆に問題事象 S を持つ人が 1 人だったとすると $p(S|C) = 1/k$ だから $p(S|C)$ は以下の範囲になる．

$$1/k \le p(S|C) \le (k - l - 1)/k \qquad (13)$$

濡れ衣被疑の確率を小さくするためには $p(S|C)$ を小さくすることがのぞましく，そのためには l は大きいほどよい．ところで，k 人のグループが全員，問題事象 S の当事者であるなら濡れ衣は発生しない．したがって，濡れ衣の観点から言えば，$p(S|C) = 1$ にするか出来るだけ小さくするかのいずれかが望ましい．

l-多様性をさらに進めた方法と言われる t-近接性[1]は疑似識別フィールドの組の値が同一の k 人の問題事象フィールドの値の分布が母集団の分布 $p(S)$ に最も近いようにグループを形成する．この場合は $p(S|C) = p(S)$ として扱うことになる．この場合は，理想的には k を十分大きくして，k 人内の分布が母集団の分布 $p(S)$ に近い必要がある．しかし，稀だが致命的な感染症など母集団の確率が非常に小さい場合は，k を極端の大きくしなければならないため，問題の解決にはほど遠い．

4. 最適化問題としての定式化

この節では，２節で述べた濡れ衣被疑の構造の定式化と 3 節で述べた k-匿名化における濡れ衣被疑とその低減方法を，最適化問題として定式化する．

k-匿名化の安全性は k だけで表されているとする．また k 人中，事象 S の当事者を s 人とし，この状態を情報環境 C とする．以下では s の値はデータベースを分析して知られているとしている．

4.1 濡れ衣被害総額

個人 j が事象 S の当事者であることを疑われる，すなわち濡れ衣被疑を受けた場合の損害額を $B_f(j,S)$ とする．例えば，1.1 節の Google サジェストによる被害では就職活動の支障や精神的被害を含め損害額を 200 万円としている．ただし，損害額は個人差が大きい．政治家や芸能人のような知名度の高い個人の場合と，無名の一般人の場合は，同じ事象で濡れ衣被疑を受けたとしても被害額に相当な差がでる．たとえば，病院への通院は政治家や芸能人の場合は致命傷になりかねないが，一般人では就活時期を除いては大きな被害はないかもしれない．しかし，個人事情まで考慮することは現状では不可能なので，損害額は個人によらない $B_f(S)$ とする．定式化は以下のようになる．

まず，k-匿名化の領域内の事象 S に関係しない k-s 人の 1 人当たりの濡れ衣被害額の期待値 $\mathrm{E}[D_f(X,S)]$ は次式となる．

$$\mathrm{E}[D_f(X,S)] = B_f(S)p_{sub}(\mathrm{f(t}(X,S),C))p(\neg\mathrm{t}(X,S)) \qquad (14)$$

ここで，$p(\neg\mathrm{t}(X,S))$ は X が事象 S に関係しない確率だから

$$p(\neg\mathrm{t}(X,S)) = 1 - s/k \qquad (15)$$

よって，式(11)は次のように書ける．

$$\mathrm{E}[D_f(X,S)] = B_f(S)p_{sub}(\mathrm{f(t}(X,S),C))(1 - s/k) \qquad (16)$$

事象 S に関係しない k-s 人が受ける濡れ衣被害総額の期待値 $\mathrm{E}[D_{k-s,f}(X,S)]$ は次の式のようになる．

$$\begin{aligned}\mathrm{E}[D_{k-s,f}(X,S)] &= (k-s)\mathrm{E}[D_f(X,S)] \\ &= (k-s)B_f(S)p_{sub}(\mathrm{f(t}(X,S),C))(1-s/k) \\ &= B_f(S)\tfrac{(k-s)^2}{k}p_{sub}(\mathrm{f(t}(X,S),C)) \qquad (17)\end{aligned}$$

次に事象 S の起こった場所を図 2 のように m 分割しよう．すると，s は s/m になり，期待値は事象 S を含む m 個の領域全てで計算しなければならない．よって，事象 S に関連する領域全部における濡れ衣被害総額の期待値 $\mathrm{E}[D_{\mathrm{m},k-s,f}(X)]$は次式となる．ただし，主観確率も S の発生確率が低い場合に変更する．これは情報環境 C が変更されたと考えられるので，これを$p_{sub}\left(\mathrm{f}(\mathrm{t}(X,S),\mathrm{C}(s/m))\right)$と書く．

$$\mathrm{E}[D_{\mathrm{m},k-s,f}(X)] = E_{m\text{領域}}\left[(k-\mathrm{s}/m)B_f(S)p_{sub}\left(\mathrm{f}(\mathrm{t}(X,S),\mathrm{C}(s/m))\right)\left(1-\frac{s}{mk}\right)\right] \quad (18)$$

4.2 事象 S の当事者が暴露された場合の損害

ある個人が事象 S の真の当事者であることが暴露される確率は s/k である．事象 S の当事者が，自分が当事者であることを疑われた場合の損害額$B_t(S)$とする．よって当事者 s 人が疑われた上で暴露されたことによる損害の期待値は次式となる．

$$E[D(X)] = B_{\mathrm{t}}(S)p_{sub}\left(\mathrm{f}(\mathrm{t}(X,S),C)\right)\frac{s}{k} \quad (19)$$

したがって，s 人の被害総額の期待値は

$$\mathrm{E}[D_s(X)] = B_{\mathrm{t}}(S)\frac{s^2}{k}p_{sub}\left(\mathrm{f}(\mathrm{t}(X,S),C)\right) \quad (20)$$

さらに m 分割した場合は次式となる．

$$\mathrm{E}[D_{\mathrm{m},s}(X)] = E_{m\text{領域}}\left[B_f(S)\left(\frac{s}{mk}\right)^2 p_{sub}\left(\mathrm{f}(\mathrm{t}(X,S),\mathrm{C}(s/m))\right)\right] \quad (21)$$

なお、同じ事象についての損害だが，濡れ衣による損害$B_f(S)$は疑われたことによる損害であり，当事者であることの暴露による損害$B_t(S)$は真実の暴露であるので，一般的には異なる．

4.3 k-匿名化による情報損失

次に k-匿名化による情報損失を算定する．これは，事象 S の当事者の推定では本来，確率 1 で推定できたものが s/k でしか推定できないので情報量の損失（つまり曖昧さの増加）1 人あたり$\log(s/k)$である．また，事象 S の当事者でない人も同様に考えれば 1 人あたり$\log(1-s/k)$である．情報損失を経済的価値（平たく言えば金額）に換算する比例定数をC_{econ}とする． すると，k-匿名化による情報損失は次式で表され，できるだけ小さくしたい．

$$C_{econ}k\left(\frac{s}{k}\cdot log\left(\frac{s}{k}\right)+\left(1-\frac{s}{k}\right)log\left(1-\frac{s}{k}\right)\right) \quad (22)$$

4.4 k-匿名化による場所情報の誤差

k-匿名化された領域の場所は、領域の中心点などが相当すると考えられる。すると、各個人 i の滞在場所は中心点からの距離だけ誤差 gerr(i)を持つ。この誤差が引き起こすデータマイニングにおける損失額を Cerr とする。

式(22)の情報損失と滞在場所誤差から生じる損失 Cerr の和がデータマイニングにおける損失となり、以下では C_{dm} と記す。よって最適化問題は k が予め与えられているなら，領域の構造によって決まるs/mを動かして解く最少化問題になる．

$$minimize_{s(\text{領域構造})}\left[\mathrm{E}[D_{\mathrm{m},k-s,f}(X)]+\mathrm{E}[D_{\mathrm{m},s}(X)]+C_{dm}\right]$$
$$\text{subject to}\quad \text{3.3 節で述べた地理的制約条件} \quad (23)$$

次に l-多様性と t-近接性を考慮した場合の最適化問題を考える．3.4 節の議論により，$p(S|C)$を直接使いたい．式(23)内部に現れる s/mk を$p\left(S|C(s/m)\right)$と変更すればよい．t-近接性を実現すると$p\left(S|C(s/m)\right)=p(S)$なので，これを使って式(23)を書き換えればよい．

これらの最適化問題は k 人のグループを同等に扱っているが，別の事象 S'が関連する隣接する領域群を考慮して最適化する必要がある大規模データに対して，これらの領域の構造を変化させて解く最適化問題の近似解法アルゴリズムの開発は今後の課題である．

5. 今後の課題

通院のような問題事象の被疑の主観確率を予測被害額と対応する対策費用の観点から導いた．次に，個人の位置情報データを対象にした k-匿名化などが行われる際に濡れ衣被疑の損害額を最少化する最適問題を定式化した．今後の課題は，4 章で定式化した最適化問題を多数の人々の位置情報履歴データに適用する場合の最適化アルゴリズムを開発である．さらに最適化問題のシミュレーションによる評価が必要である．実応用を考えると，具体的な事象 S に対する濡れ衣被疑の主観確率の推定，濡れ衣被害額の推定などが必要になる．

参考文献

1) Ninghui Li,Tiancheng Li,Venkatasubramanian, S. “t-Closeness: Privacy Beyond k-Anonymity and –Diversity”. ICDE2007, pp.106-115, 2007.
2) Machanavajjhala,A. Kifer,D. Gehrke,J. and Venkitasubramaniam, U. l-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3,2007
3) Sweeney, L. k-Anonymity: Achieving k-Anonymity Privacy Protection using Generalization and Suppresion. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, pp.571-588. 2002.
4) Yoshida M., Masaki Ikeda,M. Ono S., Sato I. and Nakagawa H. “Person Name Disambiguation by Bootstrapping”. The 33rd Annual ACM SIGIR Conference. pp.10-19, Geneva, Swiss. July 19-23, 2010
5) ショーンベルガー&クキエ(斉藤栄一郎訳). ビッグデータの正体. 講談社.2013. (V.M.Shönberger and K. Cukier. BIG DATA A Revolution That Will Transform How We Live,Work, and Think.Houghton Mifflin Harcourt Publishing Co. 2013.)
6) [閣議 2013]世界最先端 IT 国家創造宣言について．平成 25 年 6 月 14 日閣議決定. 2013