L. Sweeney. k-anonimato: um modelo para proteger a privacidade. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002; 557-570.

k-ANONIMATO: UM MODELO PARA PROTEGER A PRIVACIDADE1

LATÂNIA SWEENEY

Escola de Ciência da Computação, Carnegie Mellon University, Pittsburgh, Pensilvânia, EUA E-mail: latanya@cs.cmu.edu

Recebido em maio de 2002

Considere um detentor de dados, como um hospital ou um banco, que possui uma coleção privada de dados estruturados de campos específicos de pessoas. Suponha que o detentor dos dados queira compartilhar uma versão dos dados com os pesquisadores. Como um titular de dados pode divulgar uma versão de seu

dados privados com garantias científicas de que os indivíduos que são os titulares dos dados não podem ser reidentificados enquanto os dados permanecem úteis na prática? A solução fornecida neste documento inclui um modelo de proteção formal chamado k-anonymity e um conjunto de políticas de acompanhamento para implantação. Um comunicado fornece proteção de k-anonimato se as informações de cada pessoa contidas no comunicado não puderem ser distinguidas de pelo menos k-1 indivíduos cujas informações também aparecem no comunicado. Este artigo também examina ataques de reidentificação que podem ser realizados em lançamentos que aderem a k

anonimato, a menos que as políticas de acompanhamento sejam respeitadas. O modelo de proteção de k-anonimato é importante porque forma a base sobre a qual os sistemas do mundo real conhecidos como Datafly, µ-Argus e k-Similar fornecem garantias de proteção de privacidade.

Palavras-chave: anonimato de dados, privacidade de dados, reidentificação, fusão de dados, privacidade.

1. Introdução

A sociedade está experimentando um crescimento exponencial no número e na variedade de coleções de dados contendo informações específicas de cada pessoa, à medida que a tecnologia de computador, a conectividade de rede e o espaço de armazenamento em disco se tornam cada vez mais acessíveis. Os titulares de dados, operando de forma autônoma e com conhecimento limitado, ficam com a dificuldade de divulgar informações que não comprometam a privacidade, confidencialidade ou interesses nacionais. Em muitos casos, a própria sobrevivência do banco de dados depende da capacidade do detentor dos dados de produzir dados anônimos porque não divulgar essas informações pode diminuir a necessidade dos dados, enquanto, por outro lado, deixar de fornecer proteção adequada em uma liberação pode criar circunstâncias que prejudiquem o público ou outros.

[1] Este artigo altera e expande significativamente o documento anterior “Protegendo a privacidade ao divulgar informações: k-anonimato e sua aplicação por meio de generalização e supressão” (com Samarati) submetido ao IEEE Security and Privacy 1998 e estende partes do meu doutorado. tese “Computational Disclosure Control: A primer on data privacy protection” no Massachusetts Institute of Technology 2001.

Portanto, uma prática comum é que as organizações liberem e recebam dados específicos de pessoas com todos os identificadores explícitos, como nome, endereço e número de telefone, removidos na suposição de que o anonimato é mantido porque os dados resultantes parecem anônimos. No entanto, na maioria desses casos, os dados restantes podem ser usados para reidentificar indivíduos, vinculando ou combinando os dados com outros dados ou observando características únicas encontradas nos dados divulgados.

Em um trabalho anterior, experimentos usando dados resumidos do Censo dos EUA de 1990 foram conduzidos para determinar quantos indivíduos dentro de populações geograficamente situadas tinham combinações de valores demográficos que ocorriam com pouca frequência [1]. Combinações de poucas características geralmente se combinam em populações para identificar exclusivamente ou quase exclusivamente alguns indivíduos. Por exemplo, uma descoberta nesse estudo foi que 87% (216 milhões de 248 milhões) da população nos Estados Unidos relataram características que provavelmente os tornavam únicos com base apenas em *{5 dígitos ZIP2 , gênero, data de nascimento}*. Claramente, os dados divulgados contendo tais informações sobre esses indivíduos não devem ser considerados anônimos. No entanto, dados de saúde e outros dados específicos de uma pessoa geralmente estão disponíveis publicamente neste formulário. Abaixo está uma demonstração de como esses dados podem ser reidentificados.

Exemplo 1.Reidentificação por vinculação

A Associação Nacional de Organizações de Dados de Saúde (NAHDO) informou que 37 estados nos EUA têm mandatos legislativos para coletar dados em nível hospitalar e que 17 estados começaram a coletar dados de atendimento ambulatorial de hospitais, consultórios médicos, clínicas e assim por diante [2]. O círculo mais à esquerda na Figura 1 contém um subconjunto dos campos de informações, ou *atributos,* que a NAHDO recomenda que esses estados coletem; esses atributos incluem o CEP do paciente, data de nascimento, sexo e etnia.

Em Massachusetts, a Group Insurance Commission (GIC) é responsável pela compra de seguro saúde para funcionários do estado. O GIC coletou dados específicos do paciente com quase cem atributos por encontro ao longo das linhas mostradas no círculo mais à esquerda da Figura 1 para aproximadamente 135.000 funcionários do estado e suas famílias. Como se acreditava que os dados eram anônimos, o GIC deu uma cópia dos dados aos pesquisadores e vendeu uma cópia para a indústria [3].

Por vinte dólares, comprei a lista de registro eleitoral de Cambridge, Massachusetts, e recebi as informações em dois disquetes [4]. O círculo mais à direita na Figura 1 mostra que esses dados incluíam o nome, endereço, CEP, data de nascimento e sexo de cada eleitor. Essas informações podem ser vinculadas por CEP, data de nascimento e sexo às informações médicas,

[2] Nos Estados Unidos, um código postal refere-se ao código postal atribuído pelo Serviço Postal dos EUA. Normalmente, são usados códigos postais de 5 dígitos, embora tenham sido atribuídos códigos postais de 9 dígitos. Um código de 5 dígitos são os primeiros 5 dígitos do código de 9 dígitos.

ligando diagnóstico, procedimentos e medicamentos a indivíduos particularmente nomeados.

Por exemplo, William Weld era governador de Massachusetts na época e seus registros médicos constavam dos dados do GIC. O governador Weld morava em Cambridge, Massachusetts. De acordo com a lista de eleitores de Cambridge, seis pessoas tinham sua data de nascimento específica; apenas três deles eram homens; e ele era o único em seu CEP de 5 dígitos.

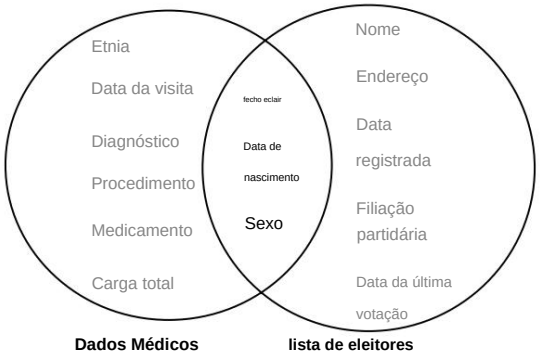

Figura 1 Vinculação para reidentificar dados

O exemplo acima fornece uma demonstração de reidentificação por vinculação direta (ou "correspondência") em atributos compartilhados. O trabalho apresentado neste artigo mostra que alterar as informações liberadas para mapear para muitas pessoas possíveis, tornando assim a ligação ambígua, pode impedir esse tipo de ataque. Quanto maior o número de candidatos fornecidos, mais ambígua a vinculação e, portanto, mais anônimos os dados.

2. Antecedentes

O problema de liberar uma versão de dados de propriedade privada para que os indivíduos que são os sujeitos dos dados não possam ser identificados não é um problema novo. Existem trabalhos existentes na comunidade de estatísticas sobre bancos de dados estatísticos e na comunidade de segurança de computadores sobre bancos de dados multiníveis a serem considerados. No entanto, nenhum desses trabalhos fornece soluções para os problemas mais amplos enfrentados no ambiente rico em dados de hoje.

2.1. Bancos de dados estatísticos

Os escritórios de estatísticas federais e estaduais em todo o mundo têm tradicionalmente sido encarregados da divulgação de informações estatísticas sobre todos os aspectos da população [5]. Mas, como outros detentores de dados, os escritórios de estatísticas também estão enfrentando uma enorme demanda por dados específicos de pessoas para aplicações como mineração de dados,

análise de custos, detecção de fraudes e pesquisa retrospectiva. Mas muitas das técnicas de banco de dados estatísticos estabelecidas, que envolvem várias maneiras de adicionar ruído [6] aos dados enquanto ainda mantêm alguma invariante estatística [7, 8], frequentemente destroem a integridade dos registros, ou *tuplas,* e assim, para muitos novos usos de dados, essas técnicas estabelecidas não são apropriadas. Willenborg e De Waal [9] fornecem uma cobertura mais extensa das técnicas estatísticas tradicionais.

2.2. Bancos de dados de vários níveis

Outra área relacionada é a agregação e inferência em bancos de dados de vários níveis [10, 11, 12, 13, 14, 15] que diz respeito à restrição da liberação de informações de classificação inferior, de modo que informações de classificação superior não possam ser derivadas. Denning e Lunt [16] descreveram um sistema de banco de dados relacional multinível (MDB) como tendo dados armazenados em diferentes classificações de segurança e usuários com diferentes autorizações de segurança.

Su e Ozsoyoglu investigaram formalmente a inferência em MDB. Eles mostraram que eliminar o compromisso de inferência precisa devido a dependências funcionais e multivaloradas é NP-completo. Por extensão a este trabalho, a eliminação precisa de todas as inferências com relação às identidades dos indivíduos cujas informações estão incluídas nos dados específicos da pessoa é normalmente impossível de garantir. Intuitivamente, isso faz sentido porque o detentor dos dados não pode considerar a priori todos os ataques possíveis. Ao tentar produzir dados anônimos, o trabalho que é objeto deste trabalho busca principalmente proteger contra ataques conhecidos. Os maiores problemas decorrem de inferências que podem ser feitas após vincular os dados divulgados a outros conhecimentos, portanto, neste trabalho, é a capacidade de vincular o resultado a fontes de dados previsíveis que deve ser controlada.

Muitos problemas de inferência de agregação podem ser resolvidos pelo design do banco de dados, mas essa solução não é prática no cenário rico em dados de hoje. No ambiente de hoje, as informações são frequentemente divididas e parcialmente replicadas entre vários detentores de dados e os detentores de dados geralmente operam de forma autônoma na tomada de decisões sobre como os dados serão divulgados. Tais decisões são normalmente tomadas localmente com conhecimento incompleto de quão sensíveis outros detentores das informações podem considerar os dados replicados. Por exemplo, quando informações um tanto antigas sobre projetos conjuntos são desclassificadas de maneira diferente pelo Departamento de Defesa e pelo Departamento de Energia, o esforço geral de desclassificação sofre; usando as duas liberações parciais, o original pode ser reconstruído em sua totalidade. Em geral, os sistemas que tentam produzir dados anônimos devem operar sem o grau de onisciência e nível de controle normalmente disponíveis no problema de agregação tradicional.

Tanto na agregação quanto no MDB, a principal técnica usada para controlar o fluxo de informações confidenciais é a *supressão,* onde as informações confidenciais e todas as informações que permitem a inferência de informações confidenciais simplesmente não são liberadas. A supressão pode reduzir drasticamente a qualidade dos dados e, no

No caso de uso estatístico, as estatísticas gerais podem ser alteradas, tornando os dados praticamente inúteis. Ao proteger os interesses nacionais, pode ser possível não liberar as informações, mas a maior demanda por dados específicos da pessoa ocorre em situações em que o detentor dos dados deve fornecer proteções adequadas, mantendo os dados úteis, como compartilhamento de dados médicos específicos da pessoa para pesquisa propósitos.

2.3. Segurança de computador não é proteção de

privacidade Uma área que pode parecer ter um ancestral comum com o assunto deste documento é o controle de acesso e a autenticação, que são áreas tradicionais associadas à segurança de computador. O trabalho nessa área garante que o destinatário da informação tenha autoridade para recebê-la. Embora as proteções de controle de acesso e autenticação possam proteger contra divulgações diretas, elas não abordam divulgações com base em inferências que podem ser extraídas dos dados divulgados.

O problema mais insidioso no trabalho que é objeto deste artigo não é tanto se o destinatário pode ter acesso ou não à informação, mas sim quais valores constituirão a informação que o destinatário receberá. Uma doutrina geral do trabalho aqui apresentado é divulgar todas as informações, mas fazê-lo de forma que as identidades das pessoas que são os sujeitos dos dados (ou outras propriedades sensíveis encontradas nos dados) sejam protegidas. Portanto, o objetivo do trabalho apresentado neste artigo está fora do trabalho tradicional de controle de acesso e autenticação.

2.4. Múltiplas consultas podem vazar inferência

Denning [17] e outros [18, 19] foram os primeiros a explorar inferências realizadas a partir de múltiplas consultas a um banco de dados. Por exemplo, considere uma tabela contendo apenas (médico, paciente, medicamento). Uma consulta listando os pacientes atendidos por cada médico, ou seja, uma relação R(médico, paciente), pode não ser sensível.
Da mesma forma, uma consulta discriminando medicamentos prescritos por cada médico também pode não ser sensível. Mas a consulta que associa os pacientes aos medicamentos prescritos pode ser delicada porque os medicamentos geralmente se correlacionam com doenças. Uma solução comum, chamada restrição de consulta, proíbe consultas que possam revelar informações confidenciais. Isso é efetivamente realizado suprimindo todas as inferências a dados confidenciais. Em contraste, este trabalho apresenta uma solução em tempo real para esse problema, defendendo que os dados sejam primeiro tornados suficientemente anônimos e, em seguida, os dados resultantes usados como base para o processamento das consultas. Fazer isso normalmente retém muito mais utilidade nos dados porque a versão resultante geralmente é menos distorcida.

Em resumo, o aumento dramático na disponibilidade de dados específicos de pessoas de detentores de dados autônomos expandiu o escopo e a natureza dos problemas de controle de inferência e exasperou as práticas operacionais estabelecidas. O objetivo deste trabalho

é fornecer um modelo para entender, avaliar e construir sistemas computacionais que controlam inferências nesse cenário.

3. Métodos

O objetivo desta seção é fornecer uma estrutura formal para construir e avaliar algoritmos e sistemas que liberam informações de forma que as informações liberadas limitem o que pode ser revelado sobre as propriedades das entidades que devem ser protegidas. Por conveniência, concentro-me em dados específicos de pessoas, de modo que as entidades são pessoas e a propriedade a ser protegida é a identidade dos sujeitos cujas informações estão contidas nos dados. No entanto, outras propriedades também podem ser protegidas. Os métodos formais fornecidos neste artigo incluem o modelo de proteção de k-anonimato. Os sistemas do mundo real Datafly [20], µ-Argus [21] ek Similar [22] motivam essa abordagem.

Salvo indicação em contrário, o termo *dados* refere-se a informações específicas de uma pessoa que são conceitualmente organizadas como uma tabela de linhas (ou registros) e colunas (ou campos). Cada linha é chamada de *tupla.* Uma tupla contém um relacionamento entre o conjunto de valores associados a uma pessoa. As tuplas dentro de uma tabela não são necessariamente únicas. Cada coluna é chamada de *atributo* e denota um campo ou categoria semântica de informação que é um conjunto de valores possíveis; portanto, um atributo também é um domínio. Os atributos dentro de uma tabela são exclusivos. Assim, observando uma tabela, cada linha é uma n-upla ordenada de valores <*d1*,*d2*,…,*dn*> tal que cada valor *dj* está no domínio da j-ésima coluna, para j=1, 2, …, *n* onde *n* é o número de atributos. Na relação correspondente a esta apresentação tabular, a única diferença é a ausência de nomes de colunas. Ullman fornece uma discussão detalhada dos conceitos de banco de dados relacional [23].

Definição 1. Atributos

Seja B(A1,…,An) uma *tabela* com um número finito de tuplas. O conjunto finito de *atributos* de B são {A1,…,An}.

Dada uma tabela B(A1,…,An), {Ai,…,Aj} ÿ {A1,…,An}, e uma tupla tÿB, eu uso t[Ai,…,Aj] para denotar a sequência de os valores, vi,…,vj, de Ai,…,Aj em t. Utilizo B[Ai,…,Aj] para denotar a projeção, mantendo tuplas duplicadas, dos atributos Ai,…Aj em B.

No restante deste trabalho, cada tupla é considerada específica de uma pessoa e duas tuplas não pertencem à mesma pessoa. Essa suposição simplifica a discussão sem perda de aplicabilidade.

Tirar uma *inferência* é passar a acreditar em um fato novo com base em outras informações. Uma *divulgação* significa que informações explícitas ou inferidas sobre uma pessoa foram divulgadas sem intenção. Esta definição pode não ser consistente com o uso coloquial, mas é usada neste trabalho de forma consistente com seu significado em

controle de divulgação estatística. Assim, *o controle de divulgação* tenta identificar e limitar as divulgações nos dados divulgados. Normalmente, o objetivo do controle de divulgação com relação a dados específicos de pessoas é garantir que os dados divulgados sejam suficientemente anônimos.

Deixe-me ser mais específico sobre como as propriedades são selecionadas e controladas. Lembre-se do exemplo de vinculação mostrado na Figura 1. Nesse caso, a necessidade de proteção centrou-se na limitação da capacidade de vincular informações liberadas a outras coleções externas. Assim, as propriedades a serem controladas são realizadas operacionalmente como atributos na coleção privada. Espera-se que o titular dos dados identifique todos os atributos nas informações privadas que possam ser usados para vinculação com informações externas. Esses atributos não incluem apenas identificadores explícitos, como nome, endereço e número de telefone, mas também incluem atributos que, combinados, podem identificar indivíduos de maneira única, como data de nascimento e sexo. O conjunto de tais atributos foi chamado de *quase-identificador* por Dalenius [24]. Portanto, operacionalmente, um objetivo deste trabalho é liberar dados específicos de pessoas de forma que a capacidade de vincular a outras informações usando o quase-identificador seja limitada.

Definição 2. Quase-identificador

Dada uma população de entidades U, uma tabela específica de entidade T(A1,…,An), f_{FC}: U ÿ : T e atributos {Ai, …, Aj} ÿ {A1, …, An} onde ∀pi∀U, tal que $f_{FC}(pi)[QT] = pi$ [T] ÿ U. Um quase-identificador de T, escrito *QT*, é um conjunto de

Exemplo 2. Quase-identificador

Seja V a tabela específica do eleitor descrita anteriormente na Figura 1 como a lista de votantes. Um quase-identificador para V, escrito *QV,* é *{nome, endereço, CEP, data de nascimento, gênero}.*

A vinculação da lista de eleitores aos dados médicos, conforme mostrado na Figura 1, demonstra claramente que *{data de nascimento, CEP, sexo}* ÿ *QV.* No entanto, *{nome, endereço}* ÿ *QV* porque esses atributos também podem aparecer em informações externas e serem usados para vinculação.

No caso do anonimato, geralmente são dados publicamente disponíveis cuja vinculação deve ser proibida e, portanto, atributos que aparecem em dados privados e também aparecem em dados públicos são candidatos a vinculação; portanto, esses atributos constituem o quase-identificador e a divulgação desses atributos deve ser controlada. Acredita-se que esses atributos possam ser facilmente identificados pelo titular dos dados.

Assunção (quase-identificador).
O titular dos dados pode identificar atributos em seus dados privados que também podem aparecer em informações externas e, portanto, pode identificar com precisão quase-identificadores.

Considere uma instância em que essa suposição está incorreta; ou seja, o titular dos dados avalia erroneamente quais atributos são sensíveis para vinculação. Nesse caso, os dados divulgados podem ser menos anônimos do que o exigido e, como resultado, os indivíduos podem ser mais facilmente identificados. Claramente, este risco não pode ser perfeitamente resolvido pelo titular dos dados porque o titular dos dados nem sempre pode saber o que cada destinatário dos dados sabe, mas políticas e contratos, que estão fora dos algoritmos, podem ajudar. Além disso, o titular dos dados pode achar necessário liberar dados que são apenas parcialmente anônimos. Novamente, políticas, leis e contratos podem fornecer proteções complementares. No restante deste trabalho, presumo que um quase-identificador adequado tenha sido reconhecido.

Como um aparte, há muitas maneiras de expandir a noção de quase-identificador para fornecer mais flexibilidade e granularidade. Ambos os sistemas Datafly e µ-Argus ponderam os atributos do quase-identificador. Para simplificar este trabalho, no entanto, considero um único quase-identificador baseado em atributos, sem pesos, aparecendo juntos em uma tabela externa ou em uma possível junção de tabelas externas.

3.1. O modelo de proteção do k-anonimato

Em um trabalho anterior, introduzi modelos básicos de proteção denominados *mapa nulo, mapa k e mapa* errado , que fornecem proteção ao garantir que a informação liberada seja mapeada para no, *k* ou entidades incorretas, respectivamente [25]. Para determinar quantos indivíduos cada tupla liberada realmente corresponde, é necessário combinar os dados liberados com dados disponíveis externamente e analisar outros possíveis ataques. Fazer tal determinação diretamente pode ser uma tarefa extremamente difícil para o titular dos dados que divulga informações. Embora eu possa assumir que o titular dos dados sabe quais dados em PT também aparecem externamente e, portanto, o que constitui um quase-identificador, os valores específicos contidos nos dados externos não podem ser assumidos. Eu, portanto, procuro proteger as informações neste trabalho satisfazendo uma restrição ligeiramente diferente nos dados divulgados, denominado requisito *de k-anonimato .* Este é um caso especial de k proteção de mapa onde *k* é aplicado nos dados liberados.

Definição 3. k-anonimato

Seja RT(A1,...,An) uma tabela e *QIRT* o quase-identificador associado a ela.
Diz-se que RT satisfaz k-anonimato se e somente se cada sequência de valores em *RT[QIRT]* aparece com pelo menos *k* ocorrências em *RT[QIRT].*

L. Sweeney. k-anonimato: um modelo para proteger a privacidade. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002; 557-570.

	Corrida	Problema	ZIP de gênero de nascimento		
t1	preto	1965	m	0214*	respiração curta
t2	preto	1965	m	0214*	dor no peito
t3	preto	1965	f	0213*	hipertensão
t4	preto	1965	f	0213*	obesidade
t5	Preto	1964	f	0213*	dor no peito
t6	preto	1964	f	0213*	obesidade
t7	branco	1964	m	0213*	dor no peito
t8	branco	1964	m	0213*	dor no peito
t9	Branco	1964			
t10	Branco	1967			
t11	Branco	1967			

Figura 2 Exemplo de k-anonimato, onde k=2 e *QI={Raça, Nascimento, Sexo, CEP}*

Exemplo 3. Tabela que adere ao k-anonimato A

Figura 2 fornece um exemplo de uma tabela T que adere ao k-anonimato. O quase-identificador para a tabela é QIT= *{Raça, Nascimento, Sexo, CEP}* ek =2. Portanto, para cada uma das tuplas contidas na tabela T, os valores da tupla que compõem o quase-identificador aparecem pelo menos duas vezes em T. Ou seja, para cada sequência de valores em *T[QIT]* existem pelo menos *2* ocorrências desses valores em *T[QIT].* Em particular, t1[QIT] = t2[QIT], t3[QIT] = t4[QIT], t5[QIT] = t6[QIT], t7[QIT] = t8[QIT] = t9[QIT] e t10 [QIT] = t11[QIT].

Lema.

Seja RT(A1,...,An) uma tabela, *QIRT =(Ai,…, Aj)* o quase-identificador associado a RT, Ai,…,Aj ÿ A1,…,An, e RT satisfaça k- anonimato. Então, cada sequência de valores em *RT[Ax]* aparece com pelo menos *k* ocorrências em *RT[QIRT]* para x=i,…,j.

Exemplo 4. *k* ocorrências de cada valor sob k-anonimato Tabela

T na Figura 2 adere ao k-anonimato, onde QIT= *{Raça, Nascimento, Sexo, CEP}* ek =2. Portanto, cada valor que aparece em um valor associado a um atributo de QI em T aparece pelo menos *k* vezes. |*T[Raça* ="negra"]| = 6. |*T[Raça* ="branca"]| = 5. |*T[Nascimento* ="1964"]| = 5. |*T[Nascimento* ="1965"]| = 4. |*T[Nascimento* ="1967"]| = 2. |*T[Gênero* ="m"]| = 6. |*T[Gênero* ="f"]| = 5. |*T[ZIP* ="0213*"]| = 9. E, |*T[ZIP* ="0214*"]| = 2.

Pode-se provar trivialmente que se os dados liberados RT satisfizerem o k-anonimato em relação ao quase-identificador QIPT, então a combinação dos dados liberados RT e as fontes externas nas quais o QIPT foi baseado, não pode se conectar ao QIPT ou a um subconjunto de seus atributos para corresponder a menos de *k* indivíduos. Esta propriedade é válida desde que todos os atributos na tabela liberada RT que estejam externamente disponíveis em

combinação (ou seja, aparecendo juntos em uma tabela externa ou em uma possível junção de tabelas externas) são definidos no quasi-identificador QIPT associado à tabela privada PT. Esta propriedade não garante que os indivíduos não possam ser identificados no RT; pode haver outros ataques de inferência que possam revelar as identidades dos indivíduos contidos nos dados. No entanto, a propriedade protege RT contra inferência de vinculação (por correspondência direta) a fontes externas conhecidas; e, neste contexto, a solução pode fornecer uma proteção eficaz contra a reidentificação de indivíduos.

CEP da corrida
Asiático 02138
Asiático 02139
Asiático 02141
Asiático 02142
Preto 02138
Preto 02139
Preto 02141
Preto 02142
Branco 02138
Branco 02139
Branco 02141
Branco 02142

PT

CEP da corrida
Pessoa 02138
Pessoa 02139
Pessoa 02141
Pessoa 02142
Pessoa 02138
Pessoa 02139
Pessoa 02141
Pessoa 02142
Pessoa 02138
Pessoa 02139
Pessoa 02141
Pessoa 02142

GT1

CEP da corrida
Asiático 02130
Asiático 02130
Asiático 02140
Asiático 02140
Preto 02130
Preto 02130
Preto 02140
Preto 02140
Branco 02130
Branco 02130
Branco 02140
Branco 02140

GT2

Figura 3 Exemplos de tabelas de k-anonimato baseadas em PT

4. Ataques contra k-anonimato

Mesmo quando é tomado cuidado suficiente para identificar o quase-identificador, uma solução que adere ao k-anonimato ainda pode ser vulnerável a ataques. Três são descritos abaixo. Felizmente, os ataques apresentados podem ser frustrados pela devida diligência de algumas práticas associadas, que também são descritas abaixo.

4.1. Ataque de correspondência não classificado contra k-anonimato

Esse ataque é baseado na ordem em que as tuplas aparecem na tabela liberada.
Embora eu tenha mantido o uso de um modelo relacional nesta discussão e, portanto, a ordem das tuplas não pode ser assumida, no uso do mundo real isso costuma ser um problema. Isso pode ser corrigido, é claro, classificando aleatoriamente as tuplas da tabela de soluções.
Caso contrário, a liberação de uma tabela relacionada pode vazar informações confidenciais.

Exemplo 5. Ataque de correspondência não classificado

As tabelas GT1 e GT2 na Figura 3 são baseadas em PT e aderem ao k-anonimato, onde QIPT = *{Race, ZIP}* ek =2. As posições das tuplas em cada tabela correspondem às de PT. Se GT1 for lançado e um lançamento subseqüente de GT2 for executado, então a correspondência direta de tuplas nas tabelas com base

na posição da tupla nas tabelas revela informações confidenciais. Por outro lado, se as posições das tuplas dentro de cada tabela forem determinadas aleatoriamente, ambas as tabelas podem ser liberadas.

4.2. Ataque de liberação complementar contra k-anonimato No

exemplo anterior, todos os atributos estavam no quasi-identificador. Isso normalmente não é o caso. É mais comum que os atributos que constituem o quase-identificador sejam eles próprios um subconjunto dos atributos liberados. Como resultado, quando uma tabela T, que adere ao k-anonimato, é liberada, ela deve ser considerada como uma junção de outras informações externas. Portanto, liberações subsequentes das mesmas informações privadas devem considerar todos os atributos liberados de T um quase identificador para proibir a vinculação em T, a menos, é claro, que liberações subsequentes sejam baseadas em T.

Exemplo 6. Ataque de liberação complementar

Considere a tabela privada PT na Figura 4. As tabelas GT1 e GT3 na Figura 5 são baseadas em PT e aderem ao k-anonimato, onde k=2 e o quasi identificador QIPT={Raça, *DataNascimento, Sexo , CEP}.* Suponha que a tabela GT1 seja liberada. Se GT3 subsequentemente também for liberado, a proteção de k-anonimato não será mais mantida, mesmo que as posições de tupla sejam determinadas aleatoriamente em ambas as tabelas. Vincular GT1 e GT3 em *{Problem}* revela a tabela LT mostrada na Figura 4. Observe como [white, 1964, male, 02138] e [white, 1965, female, 02139] são únicos em LT e, portanto, LT não satisfaz o requisito de k-anonimato imposto por GT1 e GT3. Este problema não existiria se GT3 usasse o quasi=identificador QI ÿ *{Problema}* ou se GT1 tivesse sido a base de GT3. Neste último caso, nenhum valor mais específico do que aparece em GT1 seria liberado posteriormente.

Raça	**Data de Nascimento**	**Sexo**	**ZIP**	**Problema**
preto	02141			
falta de ar	preto	20/02/1965	feminino	preto
24/8/1965	feminino	02138	chiado	preto
07/11/1964	feminino	02138		
feminino	02138	peito dor	preto	branco
23/10/1964	masculino	02138		
falta de ar	branco	branco	branco	branco
	15/03/1965	feminino	02139	hipertensão
13/08/1964				
	masculino	02139	obesidade	05/05/1964
masculino				
	02139	febre	13/02/1967	masculino
02138	vômito			
	21/03/1967	masculino	02138	dor nas costas

PT

Raça	**Data de nascimento**	**Sexo**	**CEP**	**Problema**
preto				
1965	preto	1965	preto	1965
preto	1963	141	falta	preto
masculino				
1964	branco	1964	branco	1965
branco	1964	02141	dorido	peito
feminino	02138	olho		
branco	1967	branco	1967	dolorido
feminino	02138	chiado	feminino	
		02138	obesidade	feminino
02138	dor no			
		peito	masculino	02138
falta	de	ar		
		feminino	02139	hipertensão
masculino				
		02139	obesidade	masculino
02139	febre			
		masculino	02138	vômito
masculino				
		02138	dor	nas
costas				

LT

Figura 4 Tabela privada PT e tabela vinculada LT

L. Sweeney. k-anonimato: um modelo para proteger a privacidade. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002; 557-570.

Raça	Data de Nascimento	Sexo	ZIP	Problema
preto	1965	preto	masculino	02141 falta de ar masculino
1965	pessoa		02141	dor no peito feminino 0213* olho
1965	pessoa		dolorido	feminino 0213* chiado feminino
1965	preto	1964	02138	obesidade feminino 02138 dor no
preto	1964		peito	masculino 0213* falta de ar feminino
branco	1964		0213*	hipertensão masculino 0213*
pessoa	1965		obesidade	masculino 0213* febre
branco	1964		masculino	02138 vômito masculino
branco	1964		02138	dor nas costas
branco	1967			
branco	1967			

GT1

Raça	Data de Nascimento	Sexo	ZIP	Problema
preto	1965	masculino	02141	falta de ar preto 1965 masculino
02141	dor no peito preto	1965	feminino	02138 chiado preto 1965 preto
1964	feminino	02138	obesidade	preto 1964 feminino 02138 dor no
02139	hipertensão branco	1960-69	humano	02139 febre branco 1960-69 masculino
branco	1960-69	humano	02139	febre branco 1960-69 masculino
02138	dor nas costas			

GT3

Figura 5 Duas tabelas de k-anonimato baseadas em PT na Figura 4 onde k=2

4.3. Ataque temporal contra k-anonimato

As coletas de dados são dinâmicas. Tuplas são adicionadas, alteradas e removidas constantemente. Como resultado, as liberações de dados generalizados ao longo do tempo podem estar sujeitas a um ataque de inferência temporal. Seja a tabela T0 a tabela privada original no tempo t=0. Suponha que uma solução de k-anonimato baseada em T0, que chamarei de tabela RT0, seja liberada. No tempo t, suponha que tuplas adicionais foram adicionadas à tabela privada T0, então vem Tt. Seja RTt uma solução de k-anonimato baseada em Tt que é liberada no tempo t. Como não há exigência de que RTt respeite RT0, vincular as tabelas RT0 e RTt pode revelar informações confidenciais e, assim, comprometer a proteção do k-anonimato. Como no exemplo anterior, para combater esse problema, deve-se considerar RT0 como uma junção de outras informações externas.
Portanto, todos os atributos de RT0 seriam considerados um quase identificador para lançamentos subsequentes ou os próprios lançamentos subsequentes seriam baseados em RT0.

Exemplo 7. Ataque temporal No tempo *t0,* suponha que a informação privada seja PT na Figura 4. Conforme declarado anteriormente, GT1 e GT3 na Figura 5 são soluções de k-anonimato baseadas em PT sobre o quase-identificador QIPT={Raça, *DataNascimento, Sexo, CEP}* onde k=2. Suponha que o GT1 seja lançado. Mais tarde , *t1,* PT torna -se PTt1, que é PT ÿ {[preto, 07/09/65, masculino, 02139, dor de cabeça], [preto, 04/11/65, masculino, 02139, erupção cutânea]}. Suponha que uma solução de k-anonimato baseada em PT seja fornecida e que seja chamada de GTt1. Suponha que esta tabela contenha GT3 na Figura 5; especificamente, GTt1 = GT3 ÿ {[preto, 1965, masculino, 02139, dor de cabeça], [preto, 1965, masculino, 02139, erupção cutânea]}. Como foi mostrado no exemplo anterior, GT1 e GT3 podem ser vinculados em *{Problem}* para revelar tuplas exclusivas sobre QIPT. Da mesma forma, GT1 e GTt1 podem ser vinculados para revelar as mesmas tuplas únicas. Uma maneira de combater esse problema é basear as soluções de k-anonimato em

GT1 ÿ (PTt1 – PT). Nesse caso, um resultado poderia ser GT1 ÿ {[preto, 1965, masculino, 02139, dor de cabeça], [preto, 1965, masculino, 02139, erupção cutânea]}, o que não compromete os valores distorcidos em GT1.

5. Observação final

Neste artigo, apresentei o modelo de proteção do k-anonimato, explorei ataques relacionados e forneci maneiras pelas quais esses ataques podem ser frustrados.

Agradecimentos

Em primeiro lugar, agradeço a Vicenc Torra e Josep Domingo por seu incentivo para escrever este artigo. Também dou crédito a Pierangela Samarati por nomear o k-anonimato e agradeço a ela por recomendar que eu usasse o material de maneira formal e me iniciar nessa direção em 1998. Finalmente, sou extremamente grato aos membros corporativos e governamentais do Laboratory for International Data Privacy na Carnegie Mellon University por me fornecer suporte e a oportunidade de trabalhar em problemas reais de anonimato de dados.

Referências

1 L. Sweeney, *Singularidade da Demografia Simples na População dos EUA,* LIDAP WP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000. Livro a ser publicado intitulado *The Identifability of Data.*

2 Associação Nacional de Organizações de Dados de Saúde, *Um Guia para Nível Estadual Atividades de coleta de dados de cuidados ambulatoriais* (Falls Church: Associação Nacional de Organizações de Dados de Saúde, outubro de 1996).

3 Testemunho da Group Insurance Commission perante o Massachusetts Health Care Comitê. Ver *Sessão do Joint Committee on Health Care, Massachusetts State Legislature,* (19 de março de 1997).

4 Banco de dados da lista de eleitores de Cambridge. *Cidade de Cambridge, Massachusetts.* Cambridge: fevereiro de 1997.

5 I. Fellegi. Sobre a questão do sigilo estatístico. *Journal of the American Statistical Association,* 1972, pp. 7-18.

6 J. Kim. Um método para limitar a divulgação de microdados com base em ruído aleatório e transformação *Proceedings of the Section on Survey Research Methods da American Statistical Association,* 370-374. 1986.

7 M. Palley e J. Siminoff. Divulgação baseada em metodologia de regressão de um banco de dados estatístico *Proceedings of the Section on Survey Research Methods da American Statistical Association* 382-387. 1986.

8 G. Duncan e R. Pearson. Melhorando o acesso aos dados enquanto protege a confidencialidade: perspectivas para o futuro. *Statistical Science,* maio, como artigo convidado com discussão. 1991.

L. Sweeney. k-anonimato: um modelo para proteger a privacidade. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002; 557-570.

---

9 L. Willenborg e T. De Waal. *Controle de Divulgação Estatística na Prática.* Springer Verlag, 1996.

10 T. Su e G. Ozsoyoglu. Controlando a inferência de FD e MVD em sistemas de banco de dados relacionais multinível. *IEEE Transactions on Knowledge and Data Engineering,* 3:474-- 485, 1991.

11 M. Morgenstern. Segurança e Inferência em banco de dados multinível e sistemas baseados em conhecimento. *Proc. da Conferência ACM SIGMOD ,* páginas 357--373, 1987.

12 T. Hink. Detecção de agregação por inferência em sistemas de gerenciamento de banco de dados. Em *Proc. do IEEE Symposium on Research in Security and Privacy,* páginas 96-107, Oakland, 1988.

13 T. Lunt. Agregação e inferência: Fatos e falácias. Em *Proc. do IEEE Symposium on Security and Privacy,* páginas 102--109, Oakland, CA, maio de 1989.

14 X. Qian, M. Stickel, P. Karp, T. Lunt e T. Garvey. Detecção e eliminação de canais de inferência em sistemas de banco de dados relacionais multinível. Em *Proc. do IEEE Symposium on Research in Security and Privacy,* páginas 196--205, 1993.

15 T. Garvey, T. Lunt e M. Stickel. Modelos de raciocínio abdutivo e aproximado para caracterização de canais de inferência. *IEEE Computer Security Foundations Workshop,* 4, 1991.

16 D. Denning e T. Lunt. Um modelo de dados relacional multinível. Em *Proc. do IEEE Symposium on Research in Security and Privacy,* páginas 220-234, Oakland, 1987.

17 D. Denning. *Criptografia e Segurança de Dados.* Addison-Wesley, 1982.

18 D. Denning, P. Denning e M. Schwartz. O rastreador: uma ameaça à segurança do banco de dados estatístico. *ACM Trans. on Database Systems,* 4(1):76--96, março de 1979.

19 G. Duncan e S. Mukherjee. Limitação da divulgação de microdados em bancos de dados estatísticos: tamanho da consulta e controle de consulta de amostra aleatória. Em *Proc. do 1991 IEEE Symposium on Research in Security and Privacy,* de 20 a 22 de maio, Oakland, Califórnia. 1991.

20 L. Sweeney. Garantindo o anonimato no compartilhamento de dados médicos, o sistema Datafly. *Proceedings, Journal of the American Medical Informatics Association.* Washington, DC: Hanley & Belfus, Inc., 1997.

21 A. Hundepool e L. Willenborg. µ- e ÿ-argus: software para controle de divulgação estatística. *Terceiro Seminário Internacional sobre Confidencialidade Estatística.* Sangrado: 1996.

22 L. Sweeney. Rumo à supressão ideal de detalhes ao divulgar informações médicas dados, o uso de análise de subcombinação. *Anais, MEDINFO 98.* Associação Internacional de Informática Médica. Seul, Coréia. Holanda do Norte, 1998.

23 J. Ullman. *Princípios de Banco de Dados e Sistemas de Base de Conhecimento.* Ciência da Computação Imprensa, Rockville, MD. 1988.

24 T. Dalenius. Encontrar uma agulha no palheiro – ou identificar um registro de censo anônimo. *Journal of Official Statistics,* 2(3):329-336, 1986.

25 L. Sweeney, *Proteção de privacidade de dados computacionais,* LIDAP-WP5. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000.
Livro a ser publicado intitulado *A Primer on Provide Privacy in Data.*