Grupul de lucru „articolul 29” pentru protecția datelor

0829/14/RO
WP 216

Avizul 05/2014 privind tehnicile de anonimizare

adoptat la 10 aprilie 2014

GRUPUL DE LUCRU PENTRU PROTECȚIA PERSOANELOR ÎN CEEA CE PRIVEȘTE PRELUCRAREA DATELOR CU CARACTER PERSONAL

instituit prin Directiva 95/46/CE a Parlamentului European și a Consiliului din 24 octombrie 1995,

având în vedere articolele 29 și 30,

având în vedere regulamentul său de procedură,

ADOPTĂ PREZENTUL AVIZ:

REZUMAT

În prezentul aviz, grupul de lucru analizează eficacitatea și limitele tehnicilor de anonimizare existente în raport cu contextul juridic al UE în materie de protecție a datelor și oferă recomandări pentru gestionarea tehnicilor respective luând în considerare riscul rezidual de identificare inerent fiecăreia dintre acestea.

Grupul de lucru recunoaște valoarea potențială a anonimizării, în special ca strategie care permite cetățenilor și societății în general să beneficieze de avantajele „datelor deschise", reducând în același timp riscurile pentru persoanele vizate. Cu toate acestea, studiile de caz și publicațiile din domeniul cercetării au arătat cât este de dificil să se creeze un set de date cu adevărat anonime păstrându-se, în același timp, cât mai multe dintre informațiile subiacente necesare pentru îndeplinirea sarcinii.

În conformitate cu Directiva 95/46/CE și cu alte instrumente juridice relevante ale UE, anonimizarea rezultă din prelucrarea datelor cu caracter personal în scopul de a se împiedica în mod ireversibil identificarea. Procedând astfel, operatorii de date ar trebui să țină cont de mai multe elemente, având în vedere toate mijloacele „care pot fi utilizate în mod rezonabil" pentru identificare (fie de către operator, fie de către oricare altă parte terță).

Anonimizarea constituie o prelucrare suplimentară a datelor cu caracter personal; ca atare, aceasta trebuie să îndeplinească cerința de compatibilitate prin luarea în considerare a temeiurilor juridice și a circumstanțelor prelucrării suplimentare. De asemenea, datele anonimizate nu intră în domeniul de aplicare a legislației privind protecția datelor, însă persoanele vizate pot fi în continuare îndreptățite la protecție în temeiul altor dispoziții (cum ar fi cele care protejează confidențialitatea comunicațiilor).

Principalele tehnici de anonimizare, și anume, randomizarea și generalizarea, sunt descrise în prezentul aviz. În special, avizul analizează adăugarea de zgomot, permutarea, confidențialitatea diferențială, agregarea, k-anonimatul, l-diversitatea și t-apropierea. Acesta explică principiile, punctele tari și punctele slabe ale acestora, precum și greșelile frecvente și limitele asociate utilizării fiecărei tehnici.

Avizul detaliază soliditatea fiecărei tehnici pe baza a trei criterii:

(i) mai este posibilă individualizarea unei persoane?

(ii) mai este posibil să se stabilească legături între înregistrările referitoare la o persoană? și

(iii) pot fi deduse informații referitoare la o persoană?

Cunoașterea principalelor puncte tari și puncte slabe ale fiecărei tehnici contribuie la alegerea modului de elaborare a unui proces adecvat de anonimizare într-un anumit context.

De asemenea, se abordează pseudonimizarea pentru a se clarifica unele capcane și concepții greșite: pseudonimizarea nu este o metodă de anonimizare. Aceasta se limitează la a reduce posibilitatea stabilirii unei legături între un set de date și identitatea originală a unei persoane vizate, fiind, prin urmare, o măsură de securitate utilă.

Avizul concluzionează că tehnicile de anonimizare pot oferi garanții privind viața privată și pot fi utilizate pentru a genera procese eficiente de anonimizare, dar numai dacă aplicarea

acestora este concepută în mod corespunzător — ceea ce înseamnă că trebuie să fie stabilite în mod clar condițiile prealabile (contextul) și obiectivul (obiectivele) procesului de anonimizare în vederea atingerii obiectivului de anonimizare, în același timp obținându-se seturi de date utile. Soluția optimă ar trebui să fie decisă de la caz la caz, eventual prin utilizarea unei combinații de tehnici diferite, ținându-se cont totodată de recomandările practice elaborate în prezentul aviz.

În sfârșit, operatorii de date ar trebui să țină cont de faptul că un set de date anonimizate poate prezenta în continuare riscuri reziduale pentru persoanele vizate. Într-adevăr, pe de o parte, anonimizarea și reidentificarea constituie domenii de cercetare active, iar noi descoperiri sunt publicate în mod regulat și, pe de altă parte, inclusiv datele anonimizate, cum ar fi statisticile, pot fi utilizate pentru îmbogățirea profilurilor existente ale persoanelor, creând astfel noi probleme legate de protecția datelor. Prin urmare, anonimizarea nu ar trebui să fie privită ca un exercițiu singular, iar riscurile prezente ar trebui să fie reevaluate periodic de către operatorii de date.

1 Introducere

În timp ce dispozitivele, senzorii și rețelele creează volume mari și noi tipuri de date, iar costul stocării datelor devine neglijabil, există un interes public din ce în ce mai mare și o cerere pentru reutilizarea acestor date. „Datele deschise" pot aduce avantaje clare pentru societate, persoane și organizații, dar numai dacă sunt respectate drepturile tuturor cetățenilor cu privire la protecția datelor lor cu caracter personal și a vieții private.

Anonimizarea poate fi o bună strategie de menținere a beneficiilor și de reducere a riscurilor. Odată ce un set de date este cu adevărat anonimizat și persoanele nu mai sunt identificabile, legislația europeană privind protecția datelor nu se mai aplică. Cu toate acestea, reiese clar din studiile de caz și din publicațiile din domeniul cercetării că obținerea unui set de date cu adevărat anonime dintr-un set amplu de date cu caracter personal, păstrându-se în același timp cât mai multe dintre informațiile subiacente necesare pentru îndeplinirea sarcinii, nu este o propunere simplă. De exemplu, un set de date considerate anonime poate fi combinat cu un alt set de date, astfel încât una sau mai multe persoane să poată fi identificate.

În prezentul aviz, grupul de lucru analizează eficacitatea și limitele tehnicilor de anonimizare existente în raport cu contextul juridic al UE în materie de protecție a datelor și formulează recomandări pentru o utilizare prudentă și responsabilă a tehnicilor respective în vederea elaborării unui proces de anonimizare.

2 Definiții și analiză juridică

2.1. Definiții în contextul juridic al UE

Directiva 95/46/CE se referă la anonimizare în considerentul 26 pentru a exclude datele anonimizate din domeniul de aplicare a legislației privind protecția datelor:

> *„întrucât principiile protecției trebuie să se aplice oricărei informații privind o persoană identificată sau identificabilă; întrucât, pentru a determina dacă o persoană este identificabilă este oportun să se ia în considerare toate mijloacele care pot fi utilizate în mod rezonabil fie de operator, fie de orice altă persoană pentru a identifica persoana vizată; întrucât principiile protecției nu se aplică datelor anonime astfel încât persoana vizată să nu mai fie identificabilă; întrucât codurile de conduită în sensul articolului 27 pot fi un instrument util pentru a furniza indicații asupra modului în care datele pot fi transformate în date anonime și stocate într-o formă în care nu mai pot permite identificarea persoanei vizate."*[1]

O lectură strictă a considerentului 26 oferă o definiție conceptuală a anonimizării. Considerentul 26 indică faptul că, pentru a anonimiza datele, acestea trebuie să fie private de un număr suficient de elemente, astfel încât persoana vizată să nu mai poată fi identificată. Mai exact, datele trebuie prelucrate astfel încât acestea să nu mai poată fi utilizate pentru a se identifica o persoană fizică prin intermediul „tuturor mijloacelor care pot fi utilizate în mod

[1] Ar trebui remarcat, de asemenea, că aceasta este și abordarea urmată în proiectul de regulament al UE privind protecția datelor, la considerentul 23: „pentru a se determina dacă o persoană este identificabilă, ar trebui să se ia în considerare toate mijloacele care pot fi utilizate în mod rezonabil fie de către operator, fie de către orice altă persoană în scopul identificării persoanei fizice respective".

rezonabil" fie de către operator, fie de către o parte terță. Un factor important îl reprezintă prelucrarea, care trebuie să fie un proces ireversibil. Directiva nu clarifică modul în care procesul de eliminare a posibilității de identificare ar trebui sau ar putea fi efectuat[2]. Accentul este pus pe rezultat: datele respective ar trebui să fie de așa natură încât să nu permită identificarea persoanei vizate prin intermediul „tuturor" mijloacelor „care pot fi utilizate" „în mod rezonabil". Se face trimitere la codurile de conduită ca instrument pentru stabilirea unor posibile mecanisme de anonimizare, precum și de păstrare într-o formă în care identificarea persoanei vizate nu mai este „posibilă". Astfel, directiva stabilește în mod clar un standard foarte ridicat.

Directiva asupra confidențialității și comunicațiilor electronice (Directiva 2002/58/CE) se referă, de asemenea, la „anonimizare" și la „date anonime" într-o foarte mare măsură în același sens. Considerentul 26 prevede că:

> „*Datele de transfer folosite pentru comercializarea serviciilor de comunicații sau pentru furnizarea de servicii suplimentare trebuie de asemenea șterse sau trecute în anonimat după prestarea serviciului.*"

Prin urmare, articolul 6 alineatul (1) prevede că:

> „*Datele de transfer referitoare la abonați și utilizatori prelucrate și stocate de către furnizorul rețelei de comunicații publice sau al serviciilor publice de comunicații electronice trebuie șterse sau trecute în anonimat de îndată ce nu mai sunt necesare în scopul transmiterii comunicației, fără a aduce atingere alineatelor (2), (3) și (5) din prezentul articol sau articolului 15 alineatul (1).*"

De asemenea, articolul 9 alineatul (1), prevede că:

> „*În cazul în care datele de localizare altele decât datele de transfer referitoare la abonați sau utilizatori ai rețelelor de comunicații publice sau ai serviciilor publice de comunicații electronice pot fi prelucrate, aceste date pot fi prelucrate doar dacă sunt anonime sau cu acordul utilizatorilor sau abonaților respectivi, în măsura și pe perioada cât sunt necesare în vederea furnizării unui serviciu suplimentar.*"

Raționamentul de bază este acela că rezultatul anonimizării ca tehnică aplicată datelor cu caracter personal ar trebui să fie, în stadiul tehnologic actual, la fel de permanent precum ștergerea, și anume, să facă imposibilă prelucrarea datelor cu caracter personal[3].

2.2. Analiza juridică

Analiza formulărilor referitoare la anonimizare din cadrul celor mai importante instrumente UE de protecție a datelor permite individualizarea a patru trăsături principale:

[2] Acest concept este dezvoltat în continuare la punctul 8 din prezentul aviz.

[3] Trebuie reamintit aici că anonimizarea este definită, de asemenea, în standarde internaționale precum ISO 29100 – ca fiind „procesul prin care informațiile identificabile în mod personal (PII) sunt modificate în mod ireversibil în așa fel încât o informație identificabilă în mod personal principală să nu mai poate fi identificată direct sau indirect, de către operatorul de PII independent sau în colaborare cu orice altă parte" (ISO 29100:2011). Ireversibilitatea modificării suferite de datele cu caracter personal pentru a nu permite identificarea directă sau indirectă este esențială, de asemenea, în cazul ISO. Din acest punct de vedere, există o convergență considerabilă cu principiile și conceptele care stau la baza Directivei 95/46/CE. Acest lucru se aplică, de asemenea, în cazul definițiilor incluse în unele legislații naționale (de exemplu, în Italia, Germania și Slovenia), unde accentul este pus pe imposibilitatea identificării și în care se face trimitere la „efortul disproporționat" pentru a identifica din nou (D, SI). Cu toate acestea, legislația franceză privind protecția datelor prevede faptul că datele rămân date cu caracter personal chiar dacă este extrem de dificil și puțin probabil să se reidentifice persoana vizată — cu alte cuvinte, nu există nicio dispoziție care să facă referire la verificarea „caracterului rezonabil".

- Anonimizarea poate fi un rezultat al prelucrării datelor cu caracter personal cu scopul de a împiedica în mod ireversibil identificarea persoanei vizate.

- Pot fi avute în vedere mai multe tehnici de anonimizare, nu există standarde prescriptive în legislația UE.

- Ar trebui să se acorde importanță elementelor contextuale: trebuie avute în vedere „toate" mijloacele „care pot fi utilizate în mod rezonabil" în vederea identificării de către operator și părți terțe, acordând o atenție deosebită elementelor care au devenit recent, în stadiul tehnologic actual, „utilizabile în mod rezonabil" (având în vedere creșterea puterii de calcul și instrumentele disponibile).

- Un factor de risc este inerent procesului de anonimizare: factorul de risc trebuie să fie avut în vedere în evaluarea validității tehnicilor de anonimizare – inclusiv posibilele utilizări ale oricăror date care sunt „anonimizate" prin intermediul unor astfel de tehnici – iar gravitatea și probabilitatea acestui risc ar trebui evaluate.

În prezentul aviz se utilizează sintagma „tehnică de anonimizare", nu „anonimat" sau „date anonime" pentru a se evidenția riscul rezidual inerent de reidentificare asociat oricărei măsuri de natură tehnic-organizațională al cărei scop este de a face ca datele să devină „anonime".

2.2.1. Legalitatea procesului de anonimizare

În primul rând, anonimizarea este o tehnică aplicată datelor cu caracter personal în vederea eliminării ireversibile a posibilității de identificare. Prin urmare, premisa este aceea că datele cu caracter personal trebuie să fi fost colectate și prelucrate în conformitate cu legislația aplicabilă privind păstrarea datelor într-un format identificabil.

În acest context, procesul de anonimizare, și anume, prelucrarea unor astfel de date cu caracter personal în scopul anonimizării acestora constituie „o prelucrare suplimentară". Ca atare, această prelucrare trebuie să respecte testul de compatibilitate, în conformitate cu orientările furnizate de grupul de lucru în Avizul său 03/2013 privind limitarea scopului[4].

Aceasta înseamnă că, în principiu, temeiul juridic al anonimizării se poate afla în oricare dintre motivele menționate la articolul 7 (inclusiv interesul legitim al operatorului de date), cu condiția să fie îndeplinite, de asemenea, cerințele privind calitatea datelor, conform articolului 6 din directivă și să se țină seama în mod corespunzător de circumstanțele specifice și de toți factorii menționați în avizul grupului de lucru în ceea ce privește limitarea scopului[5].

Pe de altă parte, ar trebui subliniate dispozițiile prevăzute la articolul 6 alineatul (1) litera (e) din Directiva 95/46/CE [precum și la articolul 6 alineatul (1) și articolul 9 alineatul (1) din Directiva asupra confidențialității și comunicațiilor electronice], întrucât acestea

[4] Avizul 03/2013 al grupului de lucru „articolul 29", disponibil la adresa: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5] Aceasta înseamnă, în special, că trebuie să fie efectuată o evaluare de fond, ținându-se seama de toate circumstanțele relevante, acordându-se o atenție deosebită următorilor factori cheie:
a) relația dintre scopurile pentru care au fost colectate datele cu caracter personal și scopurile prelucrării suplimentare;
b) contextul în care au fost colectate datele cu caracter personal și așteptările rezonabile ale persoanelor vizate cu privire la utilizarea ulterioară a acestor date;
c) natura datelor cu caracter personal și impactul prelucrării suplimentare asupra persoanelor vizate;
d) măsurile de protecție adoptate de operator pentru a asigura o prelucrare corectă și pentru a preveni orice impact nejustificat asupra persoanelor vizate.

demonstrează necesitatea de a nu se păstra datele cu caracter personal „într-o formă care să permită identificarea” pentru o perioadă mai lungă decât este necesar în scopul colectării sau al prelucrării suplimentare.

În sine, această dispoziție subliniază faptul că datele cu caracter personal ar trebui, cel puțin, să fie anonimizate „în mod implicit” (sub rezerva diferitelor dispoziții legale, cum ar fi cele menționate în Directiva asupra confidențialității și comunicațiilor electronice în ceea ce privește datele de transfer). În cazul în care operatorul de date dorește să păstreze astfel de date cu caracter personal, după ce scopul prelucrării inițiale și cel al prelucrării suplimentare au fost îndeplinite, ar trebui utilizate tehnici de anonimizare, astfel încât să se împiedice în mod ireversibil identificarea.

În consecință, grupul de lucru consideră că anonimizarea, în calitate de prelucrare suplimentară a datelor cu caracter personal, poate fi considerată compatibilă cu scopurile inițiale ale prelucrării, însă doar cu condiția ca procesul de anonimizare să fie realizat astfel încât să producă în mod fiabil informații anonimizate în sensul descris în prezentul document.

De asemenea, ar trebui subliniat faptul că anonimizarea trebuie să fie efectuată în conformitate cu constrângerile juridice reamintite de Curtea Europeană de Justiție în hotărârea sa în cauza C-553/07 (*College van burgemeester en wethouders van Rotterdam/M.E.E. Rijkeboer*), referitoare la necesitatea de a păstra datele într-un format identificabil pentru a permite, de exemplu, exercitarea drepturilor de acces de către persoanele vizate. CEJ a hotărât că: „*Articolul 12 litera (a) din Directiva [95/46], impune statelor membre obligația de a stabili un drept de acces la informațiile referitoare la destinatarii sau categoriile de destinatari ai datelor, precum și la conținutul informațiilor comunicate nu numai pentru prezent, ci și pentru trecut. Statele membre au sarcina de a stabili un termen de stocare a acestor informații, precum și un acces corelativ la acestea care să constituie un echilibru just între, pe de o parte, interesul persoanei vizate în privința protecției vieții sale private, în special prin intermediul căilor de intervenție și de atac prevăzute de Directiva 95/46, și, pe de altă parte, sarcina pe care obligația de stocare a acestor informații o reprezintă pentru operator”*.

Acest lucru este relevant, în special, în cazul în care operatorul de date se bazează pe articolul 7 litera (f) din Directiva 95/46/CE în ceea ce privește anonimizarea: interesul legitim al operatorului de date trebuie să fie întotdeauna în raport cu drepturile și libertățile fundamentale ale persoanelor vizate.

De exemplu, o anchetă derulată de autoritatea pentru protecția datelor (APD) din Țările de Jos în perioada 2012-2013, cu privire la utilizarea tehnologiilor de examinare în detaliu a pachetelor de date de către patru operatori de telefonie mobilă a indicat un temei juridic conform articolului 7 litera (f) din Directiva 95/46/CE pentru anonimizarea conținutului datelor de transfer cât mai curând posibil după colectarea acestor date. Într-adevăr, articolul 6 din Directiva asupra confidențialității și comunicațiilor electronice prevede că datele de transfer referitoare la abonați și utilizatori prelucrate și stocate de către furnizorul unei rețele de comunicații publice sau al unui serviciu de comunicații electronice accesibile publicului trebuie să fie șterse sau trecute în anonimat cât mai curând posibil. În acest caz, deoarece este permis în temeiul articolului 6 din Directiva asupra confidențialității și comunicațiilor electronice, există un temei juridic corespondent la articolul 7 din Directiva privind protecția datelor. De asemenea, acest lucru poate fi prezentat și astfel: în cazul în care un tip de prelucrare a datelor nu este permis în conformitate cu articolul 6 din Directiva asupra

confidențialității și comunicațiilor electronice, nu poate exista un temei juridic în articolul 7 din Directiva privind protecția datelor.

2.2.2. Potențialul de identificare a datelor anonimizate

Grupul de lucru a abordat în detaliu conceptul de date cu caracter personal în cadrul Avizului său 4/2007 privind datele cu caracter personal, axându-se pe elementele principale ale definiției de la articolul 2 litera (a) din Directiva 95/46/CE, inclusiv partea din definiție legată de noțiunea de „identificat sau identificabil". În acest context, grupul de lucru a concluzionat, de asemenea, că „datele anonimizate ar trebui, în consecință, să fie date anonime care s-au referit anterior la o persoană identificabilă pentru care identificarea nu mai este însă posibilă".

Prin urmare, grupul de lucru a clarificat deja că verificarea „mijloacelor ... utilizate în mod rezonabil" este sugerată în directivă drept un criteriu care trebuie aplicat pentru a se evalua dacă procesul de anonimizare este suficient de solid, și anume, dacă identificarea a devenit „în mod rezonabil" imposibilă. Contextul specific și circumstanțele unui caz particular au un impact direct asupra caracterului identificabil. În anexa tehnică la prezentul aviz este prezentată analiza impactului selectării celei mai adecvate tehnici.

Astfel cum s-a subliniat deja, cercetarea, instrumentele și puterea de calcul evoluează. În consecință, nu este nici posibilă, nici utilă furnizarea unei enumerări exhaustive a circumstanțelor în care identificarea nu mai este posibilă. Cu toate acestea, o serie de factori esențiali merită să fie luați în considerare și exemplificați.

În primul rând, se poate susține că operatorii de date ar trebui să se concentreze pe mijloacele concrete care ar fi necesare pentru a inversa tehnica de anonimizare, în special în ceea ce privește costul și cunoștințele necesare pentru a pune în aplicare mijloacele respective, precum și pe evaluarea probabilității și a gravității acestora. De exemplu, aceștia ar trebui să mențină un echilibru între efortul de anonimizare și costuri (atât în ceea ce privește timpul, cât și resursele necesare) în contextul creșterii disponibilității mijloacelor tehnice cu costuri reduse pentru identificarea persoanelor în seturile de date, al creșterii disponibilității publice a altor seturi de date (precum cele puse la dispoziție în legătură cu politicile privind „datele deschise") și al numeroaselor exemple de anonimizare incompletă ce produc efecte negative ulterioare, uneori ireparabile, asupra persoanelor vizate[6]. Ar trebui menționat faptul că riscul de identificare poate crește în timp și depinde, de asemenea, de dezvoltarea tehnologiei informației și comunicațiilor. Normele juridice, în cazul în care există, trebuie să fie formulate în mod neutru din punct de vedere tehnologic și, în mod ideal, trebuie să țină seama de modificările apărute în potențialul de dezvoltare a tehnologiei informației[7].

În al doilea rând, „mijloacele care pot fi utilizate în mod rezonabil pentru a se determina dacă o persoană este identificabilă" sunt cele utilizate „de operator, fie de orice altă persoană". Prin urmare, este esențial să se înțeleagă că, atunci când un operator de date nu elimină datele originale (identificabile) la nivel de eveniment, iar operatorul de date cedează o parte din respectivul set de date (de exemplu, după îndepărtarea sau mascarea datelor identificabile), setul de date rezultat conține în continuare date cu caracter personal. Numai în cazul în care

[6] Este interesant de observat că amendamentele Parlamentului European la proiectul de Regulament general privind protecția datelor, astfel cum a fost prezentat recent (21 octombrie 2013), menționează în mod specific la considerentul 23 că, „Pentru a determina dacă mijloacele pot fi utilizate în mod rezonabil pentru a identifica persoana respectivă, trebuie să se țină seama de toți factorii obiectivi, cum ar fi costurile și timpul necesar pentru identificare, ținând seama atât de tehnologia disponibilă la momentul prelucrării, cât și de dezvoltarea tehnologică".

[7] A se vedea Avizul nr. 4/2007 al grupului de lucru „articolul 29", p. 15.

operatorul de date ar agrega datele la un nivel la care evenimentele individuale nu mai sunt identificabile, setul de date rezultat poate fi calificat drept unul anonim. De exemplu: dacă o organizație colectează date privind deplasările individuale, modelele de călătorie individuale la nivel de eveniment s-ar califica în continuare drept date cu caracter personal pentru oricare parte, atât timp cât operatorul de date (sau oricare altă parte) mai are acces la datele brute originale, chiar dacă elementele directe de identificare au fost eliminate din setul furnizat părților terțe. Dimpotrivă, în cazul în care operatorul de date ar elimina datele brute și ar furniza doar statistici agregate către părți terțe la un nivel înalt, cum ar fi „luni, pe ruta X, sunt cu 160 % mai mulți pasageri decât marți", acestea s-ar califica drept date anonime.

O soluție eficientă de anonimizare împiedică toate părțile să individualizeze o persoană într-un set de date, să stabilească legături între două înregistrări în cadrul unui set de date (sau între două seturi de date separate) și să deducă orice informații într-un astfel de set de date. Prin urmare, în general, doar eliminarea elementelor de identificare directă nu este suficientă pentru a se garanta faptul că identificarea persoanei vizate nu mai este posibilă. Adeseori va fi necesar să se ia măsuri suplimentare pentru a se preveni identificarea, tot în funcție de contextul și de scopurile prelucrării pentru care sunt destinate datele anonimizate.

EXEMPLU:
Profilurile de date genetice sunt un exemplu de date cu caracter personal care pot fi expuse riscului de identificare, în cazul în care singura tehnică utilizată este aceea a eliminării identității donatorului, ca urmare a caracterului unic al anumitor profiluri. S-a demonstrat deja în literatura de specialitate[8] că asocierea dintre resursele genetice disponibile în mod public (de exemplu, registre genealogice, necrologuri, rezultatele interogărilor motoarelor de căutare) și metadatele referitoare la donatorii de ADN (momentul donării, vârsta, locul de reședință) pot dezvălui identitatea anumitor persoane, chiar dacă ADN-ul respectiv a fost donat „în mod anonim".

Ambele categorii de tehnici de anonimizare – randomizarea datelor și generalizarea[9] – prezintă puncte slabe; cu toate acestea, fiecare dintre ele poate fi adecvată în anumite circumstanțe și într-un anumit context pentru a atinge obiectivului dorit fără a pune în pericol viața privată a persoanelor vizate. Trebuie să fie clar faptul că „identificarea" nu înseamnă doar posibilitatea de a se extrage numele și/sau adresa unei persoane ci include, de asemenea, potențialul caracter identificabil prin individualizare, posibilitatea stabilirii de legături și deducții. De asemenea, pentru ca legislația privind protecția datelor să se aplice, nu contează care sunt intențiile operatorului de date sau ale destinatarului. Atât timp cât datele sunt identificabile, se aplică dispozițiile privind protecția datelor.

În cazul în care o parte terță prelucrează un set de date care a fost prelucrat printr-o tehnică de anonimizare (anonimizat și publicat de către operatorul de date inițial), ea poate face acest lucru în mod legal, fără a fi nevoie să țină seama de cerințele de protecție a datelor, cu condiția ca aceasta să nu poată identifica (direct sau indirect) persoanele vizate din setul de date inițial. Cu toate acestea, părțile terțe trebuie să ia în considerare toți factorii contextuali și circumstanțiali menționați mai sus (inclusiv caracteristicile specifice ale tehnicilor de anonimizare astfel cum au fost aplicate de către operatorul de date inițial) atunci când decid modul de utilizare și, în special, de combinare a unor astfel de date anonimizate pentru uzul propriu – întrucât consecințele acestora pot implica diferite tipuri de răspundere din partea lor. În cazul în care factorii și elementele respective sunt de natură să implice un risc inacceptabil

[8] A se vedea John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol 339, No. 6117 (18 January 2013), p. 262.
[9] Principalele caracteristici și diferențe dintre cele două tehnici de anonimizare sunt descrise în secțiunea 3 de mai jos („Analiză tehnică").

de identificare a persoanelor vizate, prelucrarea va intra din nou sub incidența legislației privind protecția datelor.

Enumerarea de mai sus nu urmărește în niciun caz să fie exhaustivă ci, mai degrabă, să ofere orientări generale privind modul de abordare a evaluării potențialului de identificare a unui anumit set de date care face obiectul unui proces de anonimizare conform diferitelor tehnici disponibile. Toți factorii de mai sus pot fi considerați tot atâția factori de risc care trebuie avuți în vedere atât de către operatorii de date atunci când anonimizează seturile de date, cât și de către terți atunci când utilizează seturile de date „anonimizate" în scop propriu.

2.2.3. Riscurile utilizării de date anonimizate

Atunci când iau în considerare posibilitatea utilizării tehnicilor de anonimizare, operatorii de date trebuie să țină cont de următoarele riscuri:

- O capcană specifică este aceea de a considera datele pseudonimizate ca fiind echivalente cu datele anonimizate. Secțiunea care cuprinde analiza tehnică va explica faptul că datele pseudonimizate nu pot fi echivalate cu informațiile anonimizate, întrucât acestea continuă să permită individualizarea unei persoane vizate și posibilitatea creării de legături între aceasta și diferitele seturi de date. Pseudonimatul este de natură să permită identificarea și, prin urmare, se încadrează în domeniul de aplicare a regimului juridic de protecție a datelor. Acest lucru este deosebit de relevant în contextul cercetării istorice, statistice sau științifice[10].

EXEMPLU:
Un exemplu tipic de concepții greșite cu privire la pseudonimizare este furnizat de binecunoscutul „incident AOL (America On Line). În 2006, o bază de date cuprinzând douăzeci de milioane de cuvinte cheie de căutare pentru peste 650 000 de utilizatori pentru o perioadă de 3 luni a fost făcută publică, singura măsură de protecție a vieții private constând doar în înlocuirea identificatorului de utilizator AOL cu un atribut numeric. Acest lucru a condus la identificarea publică și localizarea unora dintre utilizatori. Șirurile de interogare pseudonimizate ale motoarelor de căutare, în special în combinație cu alte atribute, cum ar fi adrese IP sau alți parametri de configurare ai clientului, dețin o capacitate foarte ridicată de identificare.

- O a doua eroare este aceea de a se considera că datele anonimizate în mod corespunzător (care au îndeplinit toate condițiile și criteriile menționate anterior și care, prin definiție, nu se încadrează în domeniul de aplicare al Directivei privind protecția datelor) privează persoanele de oricare garanție – în primul rând, pentru că alte acte legislative se pot aplica utilizării acestor date. De exemplu, articolul 5 alineatul (3) din Directiva asupra confidențialității și comunicațiilor electronice împiedică stocarea și accesul la „informații" de orice tip (inclusiv informațiile fără caracter personal) la echipamentele terminale fără consimțământul abonatului/utilizatorului deoarece aceasta face parte din principiul mai amplu al confidențialității comunicațiilor.

- O a treia neglijență ar rezulta, de asemenea, din faptul că nu se ia în considerare impactul asupra persoanelor, în anumite circumstanțe, al datelor anonimizate în mod adecvat, în special în cazul creării de profiluri. Sfera vieții private a unei persoane este protejată de articolul 8 din Convenția europeană a drepturilor omului și de articolul 7 din Carta drepturilor fundamentale a UE; prin urmare, chiar dacă este posibil ca legile privind protecția datelor să nu se mai aplice în cazul acestui tip de date, modul în care sunt utilizate seturile de date anonimizate și publicate spre a fi utilizate de către părți terțe poate determina a pierdere a confidențialității. Este necesară o prudență specială în manipularea informațiilor anonimizate, în special atunci când astfel de informații sunt utilizate (deseori în combinație cu alte date) pentru luarea unor

[10] A se vedea, de asemenea, Avizul nr. 4/2007 al grupului de lucru „articolul 29", p. 18-20.

decizii care produc efecte (chiar dacă în mod indirect) asupra persoanelor. Astfel cum s-a evidențiat deja în prezentul aviz și cum a fost clarificat de către grupul de lucru, în special în Avizul privind conceptul de „limitare a scopului" (Avizul 03/2013)[11], așteptările legitime ale persoanelor vizate referitoare la prelucrarea ulterioară a datelor lor ar trebui să fie evaluate având în vedere factorii contextuali relevanți, cum ar fi natura relației dintre persoanele vizate și operatorii de date, obligațiile legale aplicabile, transparența operațiunilor de prelucrare.

3 Analiza tehnică, soliditatea tehnologiilor și erori frecvente

Există diferite practici și tehnici de anonimizare cu grade diferite de soliditate. Această secțiune va aborda principalele puncte care trebuie avute în vedere de către operatorii de date atunci când le aplică, ținând seama, în special, de garanția pe care o poate oferi tehnica respectivă, luând în considerare stadiul tehnologic actual și având în vedere trei riscuri care sunt esențiale pentru anonimizare:

- *individualizarea*, ceea ce înseamnă posibilitatea de a se izola parțial sau integral înregistrările care duc la identificarea unei persoane în setul de date;
- *posibilitatea stabilirii de legături,* care înseamnă capacitatea de a se crea legături, cel puțin între două înregistrări privind aceeași persoană vizată sau un grup de persoane vizate (fie în aceeași bază de date, fie în două baze de date diferite). Dacă un atacator poate stabili (de exemplu, prin analiza corelației) că două înregistrări sunt atribuite aceluiași grup de persoane, dar nu poate individualiza persoane în cadrul acestui grup, tehnica prezintă rezistență împotriva „individualizării", însă nu și împotriva posibilității de a se stabili legături;
- *deducția*, care constă în posibilitatea de a se deduce, cu o probabilitate semnificativă, valoarea unui atribut din valorile unui set de alte atribute.

Astfel, o soluție împotriva acestor trei riscuri ar fi protecția solidă împotriva reidentificării efectuate prin cele mai probabile și rezonabile mijloace pe care operatorul de date și oricare parte terță le pot utiliza. Grupul de lucru subliniază, în acest sens, că tehnicile de eliminare a posibilității identificării și de anonimizare fac obiectul cercetărilor în curs, iar astfel de cercetări au arătat în mod consecvent că orice tehnică prezintă puncte slabe. În sens larg, există două abordări diferite ale anonimizării: prima se bazează pe **randomizare,** în timp ce a doua se bazează pe **generalizare**. De asemenea, avizul abordează și alte concepte, cum ar fi *pseudonimizarea*, *confidențialitatea diferențială*, *l-diversitatea*, *t-apropierea*.

Prezentul aviz folosește următoarea terminologie în cadrul acestei secțiuni: un set de date este compus din diferite înregistrări referitoare la persoane (persoanele vizate). Fiecare înregistrare se referă la o persoană vizată și este alcătuită dintr-un set de valori (sau „intrări", de exemplu: 2013) pentru fiecare atribut (de exemplu, an). Un set de date este o colecție de înregistrări care pot fi modelate în mod alternativ, sub forma unui tabel (sau a unui set de tabele) sau a unui grafic adnotat/ponderat, care este din ce în ce mai folosit în prezent. Exemplele din prezentul aviz se vor referi la tabele, însă acestea sunt aplicabile, de asemenea, altor reprezentări grafice ale înregistrărilor. Combinațiile de atribute referitoare la o persoană vizată sau un grup de persoane vizate pot fi menționate ca fiind cvasi-identificatori. În unele cazuri, un set de date poate avea înregistrări multiple privind aceeași persoană. Un „atacator" este o

[11] Disponibil la adresa http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

parte terță (și anume, nici operatorul de date, nici prelucrătorul de date) care accesează înregistrările originale fie în mod accidental, fie intenționat.

3.1. Randomizarea

Randomizarea este o categorie de tehnici care modifică veridicitatea datelor în scopul eliminării legăturii puternice dintre date și individ. Dacă datele sunt suficient de incerte, atunci acestea nu mai pot face referire la o anumită persoană. Randomizarea în sine nu va reduce unicitatea fiecărei înregistrări, întrucât fiecare înregistrare va fi în continuare derivată de la o singură persoană vizată, însă aceasta poate proteja împotriva atacurilor/riscurilor de deducție și poate fi combinată cu tehnici de generalizare pentru a furniza garanții mai solide privind protecția vieții private. Ar putea fi necesare tehnici suplimentare pentru a garanta faptul că o înregistrare nu poate conduce la identificarea unei persoane.

3.1.1. Adăugarea de zgomot

Tehnica adăugării de zgomot este deosebit de utilă atunci când atributele ar putea avea un efect negativ important asupra persoanelor și constă în modificarea atributelor din setul de date astfel încât acestea să fie mai puțin precise, păstrându-se în același timp distribuția globală. Atunci când prelucrează un set de date, un observator va considera că valorile sunt corecte, însă acest lucru va fi adevărat doar într-o anumită măsură. De exemplu, în cazul în care înălțimea unei persoane a fost măsurată inițial la cel mai apropiat centimetru, setul de date anonimizate poate conține o precizie a înălțimii de doar ± 10 cm. Dacă această tehnică este aplicată în mod eficient, o parte terță nu va putea să identifice o persoană și nici nu ar trebui să fie în măsură să repare datele sau să detecteze în alt mod maniera în care datele au fost modificate.

Adăugarea de zgomot va trebui să fie combinată, în general, cu alte tehnici de anonimizare, cum ar fi eliminarea atributelor evidente și a cvasi-identificatorilor. Nivelul de zgomot ar trebui să depindă de necesitatea nivelului de informații solicitat și de impactul asupra vieții private a persoanelor ca urmare a divulgării atributelor protejate.

3.1.1.1. Garanții

- *Individualizarea:* este posibilă în continuare identificarea înregistrărilor cu privire la o persoană (eventual într-o manieră neidentificabilă), chiar dacă înregistrările sunt mai puțin fiabile.
- *Posibilitatea stabilirii de legături:* este posibilă în continuare stabilirea de legături între înregistrările cu privire la aceeași persoană, însă înregistrările sunt mai puțin fiabile și, prin urmare, o înregistrare reală poate fi asociată cu una adăugată în mod artificial (de exemplu, cu „zgomot”). În unele cazuri, o atribuire greșită ar putea expune persoana vizată unui nivel de risc semnificativ și chiar mai ridicat decât o atribuire corectă.
- *Deducția:* atacurile prin deducție pot fi posibile, însă rata de succes va fi mai scăzută și sunt posibile unele rezultate fals pozitive (și fals negative).

3.1.1.2. Erori frecvente

- Adăugarea de zgomot incompatibil: dacă zgomotul nu este viabil din punct de vedere semantic (și anume, este „disproporționat” și nu respectă logica dintre atributele dintr-un set), atunci un atacator care are acces la baza de date va putea filtra zgomotul și, în unele cazuri, va putea regenera intrările care lipsesc. De asemenea, dacă setul de date

este prea rar[12], poate fi posibilă în continuare stabilirea de legături între intrările de date cu zgomot adăugat și o sursă externă.

- Presupunerea faptului că adăugarea de zgomot este suficientă: adăugarea de zgomot este o măsură complementară care face mai dificilă pentru atacator extragerea de date cu caracter personal. Exceptând cazul în care zgomotul este mai mare decât informațiile conținute în setul de date, nu ar trebuie să se presupună că adăugarea de zgomot reprezintă o soluție de sine stătătoare pentru anonimizare.

3.1.1.3. Limitele adăugării de zgomot

Un experiment celebru privind reidentificarea este cel efectuat cu privire la baza de date a clienților furnizorului de conținut video Netflix. Cercetătorii au analizat proprietățile geometrice ale acestei baze de date constând în peste 100 de milioane de aprecieri pe o scară de la 1 la 5 pentru un număr de peste 18 000 de filme, exprimate de aproape 500 000 de utilizatori, făcută publică de către societate, după ce a fost „anonimizată" în conformitate cu o politică internă de confidențialitate, fiind eliminate toate informațiile de identificare ale clienților cu excepția aprecierilor și a datelor. A fost adăugat zgomot în sensul că aprecierile au fost ușor majorate sau reduse.

În pofida acestui fapt, s-a constatat că 99 % din înregistrările utilizatorilor puteau fi identificate în mod unic în setul de date folosind 8 aprecieri și date cu o eroare de 14 zile ca și criterii de selecție, în timp ce diminuarea criteriilor de selecție (2 aprecieri și o eroare de 3 zile) a permis în continuare identificarea a 68 % din utilizatori[13].

3.1.2. Permutarea

Permutarea constă în amestecarea valorilor atributelor dintr-un tabel, astfel încât unele dintre acestea să fie legate în mod artificial de diferite persoane vizate; această tehnică este utilă atunci când este important să se păstreze distribuția exactă a fiecărui atribut din setul de date.

Permutarea poate fi considerată o formă specială de adăugare de zgomot. În tehnica de adăugare de zgomot clasică, atributele sunt modificate cu valori randomizate. Generarea de zgomot coerent poate fi o sarcină dificilă, iar modificarea ușoară a valorilor atributelor ar putea să nu ofere un nivel adecvat de protecție a vieții private. Ca alternativă, tehnicile de permutare modifică valorile dintr-un set de date doar prin schimbarea acestora de la o înregistrare la alta. Un astfel de transfer va garanta faptul că intervalul și distribuția valorilor vor rămâne aceleași, însă nu și corelațiile dintre valori și persoane. Dacă două sau mai multe atribute au o legătură logică sau o corelare statistică și sunt permutate în mod independent, o astfel de relație va fi distrusă. Prin urmare, ar putea fi importantă permutarea unui set de atribute legate astfel încât să nu se rupă legătura logică, în caz contrar un atacator ar putea identifica atributele permutate și ar putea inversa permutarea.

De exemplu, dacă se consideră un subset de atribute în cadrul unui set de date medicale, cum ar fi „motive pentru spitalizare/simptome/departament responsabil", o puternică relație logică va lega aceste valori în majoritatea cazurilor, iar permutarea unei singure valori ar fi astfel detectată și chiar ar putea fi inversată.

[12] Acest concept este dezvoltat în continuare în anexă, p. 30.

[13] Narayanan, A., & Shmatikov, V. (2008, mai). Robust de-anonymization of large sparse datasets, Security and Privacy, 2008. SP 2008. IEEE Symposium on Security and Privacy (p. 111-125). IEEE.

În mod similar adăugării de zgomot, permutarea în sine nu poate asigura anonimizarea și ar trebui să fie întotdeauna combinată cu eliminarea atributelor evidente/cvasi-identificatorilor.

3.1.2.1. Garanții

- *Individualizarea:* la fel ca și în cazul adăugării de zgomot, este posibilă în continuare identificarea înregistrărilor cu privire la o persoană, cu toate acestea înregistrările sunt mai puțin fiabile.
- *Posibilitatea stabilirii de legături:* dacă permutarea afectează atributele și cvasi-identificatorii, aceasta ar putea împiedica stabilirea de legături „corecte" între atribute, atât la nivel intern, cât și la nivel extern, și un set de date, însă permite în continuare posibilitatea stabilirii de legături „incorecte", întrucât o intrare reală poate fi asociată cu o altă persoană vizată.
- *Deducția:* pot fi efectuate în continuare deducții din setul de date, în special în cazul în care atributele sunt corelate sau prezintă relații logice puternice; cu toate acestea, necunoscând atributele permutate, atacatorul trebuie să ia în considerare faptul că deducția sa se bazează pe o ipoteză incorectă și, prin urmare, rămâne posibilă doar deducția probabilistică.

3.1.2.2. Erori frecvente

- Selectarea atributului greșit: permutarea atributelor nesensibile sau care nu prezintă riscuri nu ar conduce la o creștere semnificativă în termeni de protecție a datelor cu caracter personal. Într-adevăr, dacă atributele sensibile/riscante ar fi în continuare asociate cu atributul original, atunci un atacator ar fi în continuare în măsură să extragă informații sensibile referitoare la persoane.
- Permutarea aleatorie a atributelor: dacă două atribute sunt strâns corelate, atunci permutarea aleatorie a atributelor nu va oferi garanții solide. Această eroare frecventă este ilustrată în tabelul 1.
- Presupunerea faptului că permutare este suficientă: la fel ca și în cazul adăugării de zgomot, permutarea în sine nu asigură anonimatul, de aceea ar trebui combinată cu alte tehnici, cum ar fi eliminarea atributelor evidente.

3.1.2.3. Limitele permutării

Acest exemplu ilustrează modul în care atributele permutate în mod aleatoriu au drept rezultat garanții slabe în ceea ce privește protecția vieții private atunci când există legături logice între diferite atribute. În urma încercării de anonimizare, este foarte ușor să se deducă venitul fiecărei persoane în funcție de profesie (și anul nașterii). De exemplu, se poate afirma, prin analizarea directă a datelor, că directorul executiv din tabel s-a născut foarte probabil în anul 1957 și are cel mai mare salariu, în timp ce șomerul s-a născut în 1964 și are cel mai scăzut venit.

Anul	Sex	Profesie	Venit (permutat)
1957	M	Inginer	70 000
1957	M	Director executiv	5 000
1957	M	Șomer	43 000
1964	M	Inginer	100 000
1964	M	Administrator	45 000

Tabelul 1. Un exemplu ineficient de anonimizare prin permutarea atributelor corelate

3.1.3. Confidențialitate diferențială

Confidențialitatea diferențială[14] intră în categoria tehnicilor de randomizare, cu o abordare diferită: în timp ce, în fapt, inserția de zgomot intervine în prealabil, atunci când setul de date ar trebui să fie publicat, confidențialitatea diferențială poate fi folosită atunci când operatorul de date generează opinii anonimizate cu privire la un set de date, păstrând în același timp o copie a datelor originale. Astfel de opinii anonimizate ar fi generate în mod obișnuit prin intermediul unui subset de interogări pentru o anumită parte terță. Subsetul include unele elemente de zgomot aleatorii adăugat în mod deliberat *ex-post*. Confidențialitatea diferențială indică operatorului de date câte elemente de zgomot trebuie să adauge și sub ce formă pentru a obține garanțiile necesare privind protecția vieții private[15]. În acest context, va fi deosebit de important să se monitorizeze în permanență (cel puțin pentru fiecare nouă interogare), orice posibilitate de a identifica o persoană în setul de rezultate din interogare. Cu toate acestea, trebuie clarificat faptul că tehnicile de confidențialitate diferențială nu vor modifica datele originale și, prin urmare, atât timp cât datele originale rămân, operatorul de date este în măsură să identifice persoanele în rezultatele interogărilor confidențialității diferențiale, luând în considerare toate mijloacele care pot fi utilizate în mod rezonabil. Aceste rezultate trebuie să fie considerate, de asemenea, date cu caracter personal.

Un avantaj al unei abordări bazate pe confidențialitate diferențială constă în faptul că seturile de date sunt furnizate părților terțe autorizate ca răspuns la o interogare specifică, mai degrabă decât prin publicarea unui set de date individual. Pentru a se facilita efectuarea auditului, o listă a tuturor interogărilor și solicitărilor poate fi păstrată de către operatorul de date, care să garanteze faptul că părțile terțe nu au acces la date pentru care nu sunt autorizate. De asemenea, o interogare poată fi supusă tehnicilor de anonimizare, inclusiv adăugarea de zgomot sau substituție în scopul de a se proteja viața privată. Se efectuează în continuare cercetări pentru identificarea unui mecanism interactiv eficient interogare-răspuns, care, în același timp, să poată răspunde oricăror întrebări într-un mod destul de precis (și anume, într-un mod mai puțin zgomotos) și să protejeze viața privată.

Pentru limitarea atacurilor prin deducție și stabilirea de legături, este necesar să se mențină o evidență a interogărilor efectuate de o entitate și să se monitorizeze informațiile obținute cu privire la persoanele vizate; prin urmare, bazele de date privind „confidențialitatea diferențială" nu ar trebui introduse pe motoare de căutare deschise care nu oferă trasabilitatea entităților care efectuează interogări.

[14] Dwork, C. (2006). Differential privacy. *Automata, languages and programming* (pp. 1-12). Springer, Berlin Heidelberg.
[15] Cf. Ed Felten (2012). Protecting privacy by adding noise. URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

3.1.3.1 Garanții

- *Individualizarea:* în cazul în care rezultatul este reprezentat doar de statistici, iar regulile aplicate sunt bine alese, nu ar trebui să fie posibilă utilizarea răspunsurilor pentru a se identifica o persoană.
- *Posibilitatea stabilirii de legături:* prin utilizarea solicitărilor multiple, ar putea fi posibilă stabilirea unor legături între intrările referitoare la o anumită persoană între două răspunsuri.
- *Deducţia:* se pot deduce informații referitoare la persoane sau grupuri prin utilizarea solicitărilor multiple.

3.1.3.2. Erori frecvente

- Neinjectarea unui număr suficient de elemente de zgomot: pentru a se preveni stabilirea de legături cu cunoștințele generale, dificultatea constă în furnizarea de dovezi minime cu privire la faptul dacă o anumită persoană vizată sau un grup de persoane vizate a contribuit la setul de date. Principala dificultate din perspectiva protecției datelor este de a se putea genera cantitatea adecvată de zgomot care să fie adăugată la răspunsurile adevărate astfel încât să se protejeze viața privată a persoanelor, în același timp menținându-se utilitatea răspunsurilor publicate.

3.1.3.3 Limitele confidențialității diferențiale

Tratarea fiecărei interogări în mod independent: o combinare a rezultatelor interogărilor poate permite divulgarea unor informații menite să rămână secrete. În cazul în care nu este păstrat un istoric al interogărilor, un atacator poate elabora întrebări multiple către o bază de date privind „confidențialitatea diferențială” prin care să reducă, în mod progresiv, amplitudinea eșantionului rezultat până când ar putea apărea un caracter specific al unei persoane vizate individuale sau al unui grup de persoane vizate, în mod determinist sau cu o probabilitate foarte mare. De asemenea, o altă dificultate este să se evite eroarea de a considera că datele sunt anonime pentru o parte terță în timp ce operatorul de date poate identifica în continuare persoana vizată în baza de date originală, luând în considerare toate mijloacele care pot fi utilizate în mod rezonabil.

3.2. Generalizarea

Generalizarea este cea de-a doua categorie de tehnici de anonimizare. Această abordare constă în generalizarea, sau diluarea, atributelor persoanelor vizate prin modificarea dimensiunilor sau a nivelului amplorii acestora (și anume, o regiune mai degrabă decât un oraș, o lună mai degrabă decât de o săptămână). În timp ce generalizarea poate fi eficientă pentru a se preveni individualizarea, aceasta nu permite o anonimizare eficientă în toate cazurile; în special, aceasta presupune abordări cantitative specifice și sofisticate pentru a se preveni posibilitatea stabilirii de legături și deducția.

3.2.1. Agregarea și K-anonimatul

Tehnicile de agregare și de k-anonimat vizează să prevină individualizarea unei persoane vizate prin gruparea acestora cu, cel puțin, *k* alte persoane. Pentru a se realiza acest lucru, valorile atributelor sunt generalizate astfel încât fiecare persoană să partajeze aceeași valoare. De exemplu, prin diminuarea granularității unei locații de la un oraș la o țară, este inclus un număr mai mare de persoane vizate. Datele de naștere individuale pot fi generalizate în intervale de date sau grupate în funcție de lună sau an. Alte atribute numerice (de exemplu,

salarii, greutate, înălțime sau doza unui medicament) pot fi generalizate prin intervale de valori (de exemplu, salariu 20 000 EUR – 30 000 EUR). Aceste metode pot fi utilizate atunci când corelarea valorilor punctuale ale atributelor ar putea crea cvasi-identificatori.

3.2.1.1. Garanții

- *Individualizarea:* având în vedere că aceleași atribute sunt partajate acum de k utilizatori, nu ar trebui să mai fie posibilă identificarea unei persoane în cadrul unui grup de k utilizatori.
- *Posibilitatea stabilirii de legături:* cu toate că posibilitatea stabilirii de legături este limitată, este posibil în continuare să se stabilească legături între înregistrările grupurilor de k utilizatori. Ulterior, în cadrul acestui grup, probabilitatea ca două înregistrări să corespundă acelorași pseudo-identificatori este 1/k (care ar putea fi semnificativ mai mare decât probabilitatea ca înregistrările respective să nu poată fi corelate).
- *Deducția:* principalul punct slab al modelului k-anonimat este acela că nu previne orice tip de atac prin deducție. Într-adevăr, dacă toate cele k persoane se află în același grup, atunci, dacă se cunoaște grupul căruia îi aparține o persoană, este ușor să se extragă valoarea acestei proprietăți.

3.2.1.2. Erori frecvente

- Lipsa unor cvasi-identificatori: un parametru esențial atunci când se analizează posibilitatea utilizării k-anonimatului este reprezentat de pragul k. Cu cât este mai mare valoarea lui k, cu atât sunt mai solide garanțiile privind protecția vieții private. O eroare frecventă este aceea de a majora în mod artificial valoarea lui k prin reducerea setului analizat de cvasi-identificatori. Reducerea cvasi-identificatorilor facilitează crearea de grupuri de k-utilizatori datorită puterii inerente de identificare asociate altor atribute (în special în cazul în care unele dintre acestea sunt sensibile sau dețin un nivel foarte ridicat al entropiei, astfel cum este cazul atributelor foarte rare). Neluarea în considerare a tuturor cvasi-identificatorilor atunci când se selectează atributul care urmează să fie generalizat constituie o eroare esențială; dacă anumite caracteristici pot fi utilizate pentru a se individualiza o persoană într-un grup de k persoane, atunci generalizarea nu reușește să protejeze unele persoane (a se vedea exemplul din tabelul 2).
- O valoare mică pentru k: vizarea unei valori mici pentru k este, de asemenea, problematică. În cazul în care k este prea mic, ponderea fiecărei persoane în cadrul unui grup este prea importantă, iar atacurile prin deducție au o rată mai mare de succes. De exemplu, dacă k = 2, atunci probabilitatea ca două persoane să prezinte aceeași proprietate este mai mare decât în cazul în care k > 10.
- Formarea de grupuri cu persoane cu ponderi diferite: gruparea unui set de persoane cu o distribuție inegală a atributelor poate fi, de asemenea, problematică. Impactul istoricului unei persoane asupra unui set de date va varia: unele vor reprezenta o parte semnificativă pentru intrări, în timp ce contribuțiile altora vor rămâne destul de nesemnificative. Prin urmare, este important să se asigure faptul că k este suficient de mare astfel încât persoanele să nu reprezinte o pondere prea importantă din intrările din cadrul unui grup.

3.1.3.3. Limitele k-anonimatului

Principala problemă a k-anonimatului este aceea că nu împiedică atacurile prin deducție. În următorul exemplu, dacă atacatorul știe că o anumită persoană se află într-un set de date și că aceasta s-a născut în 1964, atunci acesta știe, de asemenea, că persoana a suferit un atac de cord. În plus, dacă se știe că setul de date a fost obținut de la o organizație franceză, atunci fiecare persoană locuiește în Paris, întrucât primele trei cifre ale codurilor poștale din Paris sunt750*).

Anul	Sex	Codul poștal	Diagnostic
1957	M	750*	Atac de cord
1957	M	750*	Colesterol
1957	M	750*	Colesterol
1964	M	750*	Atac de cord
1964	M	750*	Atac de cord

Tabelul 2. Exemplu de k-anonimizare concepută deficitar

3.2.2. L-diversitatea/t-apropierea

L-diversitatea extinde k-anonimatul pentru a garanta faptul că atacurile deterministe prin deducție nu mai sunt posibile prin asigurarea faptului că în fiecare clasă de echivalență fiecare atribut are cel puțin *l* valori diferite.

Un obiectiv de bază de atins este acela de a se limita apariția de clase de echivalență cu o variabilitate scăzută a atributelor, astfel încât un atacator cu cunoștințe generale cu privire la o anumită persoană vizată să rămână întotdeauna cu un nivel semnificativ de incertitudine.

L-diversitatea este utilă pentru protejarea datelor împotriva atacurilor prin deducție atunci când valorile atributelor sunt bine distribuite. Trebuie subliniat însă că această tehnică nu poate împiedica scurgerea de informații dacă atributele din cadrul unei partiții sunt distribuite în mod inegal sau aparțin unui interval redus de valori sau sensuri semantice. În cele din urmă, l-diversitatea face obiectul unor atacuri probabilistice prin deducție.

T-apropierea este o rafinare a l-diversității prin faptul că urmărește să creeze clase echivalente care se aseamănă prin distribuția inițială a atributelor în tabel. Această tehnică este utilă atunci când este important să se păstreze datele cât mai apropiate de cele originale; în acest sens, o constrângere suplimentară este introdusă în ceea ce privește clasa de echivalență și anume, aceea că ar trebui să existe nu doar cel puțin *l* valori diferite în cadrul fiecărei clase de echivalență ci, de asemenea, fiecare valoare ar trebui să fie reprezentată de câte ori este necesar pentru a reflecta distribuția inițială a fiecărui atribut.

3.2.2.1. Garanții

- *Individualizarea:* la fel ca și în cazul k-anonimatului, l-diversitatea și t-apropierea pot garanta faptul că înregistrările cu privire la o persoană nu pot fi identificate în baza de date.
- *Posibilitatea stabilirii de legături*: l-diversitatea și t-apropierea nu reprezintă o îmbunătățire față de k-anonimat în ceea ce privește imposibilitatea creării de legături. Problema este aceeași precum în cazul oricărui grup: probabilitatea ca aceleași intrări

să aparțină unei aceleiași persoane vizate este mai mare decât 1/N (unde N este numărul de persoane vizate din baza de date).

- *Deducţia:* principala îmbunătățire a l-diversității și a t-apropierii față de k-anonimat este aceea că nu mai este posibil să se producă atacuri prin deducție împotriva unei baze de date „l-diversă" sau „t apropiată" cu un nivel de încredere de 100 %.

3.2.2.2. Erori frecvente

- Protejarea valorilor atributelor sensibile prin amestecarea acestora cu alte atribute sensibile: Nu este suficient să se dețină două valori ale unui atribut în cadrul unui grup pentru a furniza garanții privind protecția vieții private. În fapt, distribuția valorilor sensibile în cadrul fiecărui grup ar trebui să se asemene cu distribuția valorilor respective în cadrul populației totale sau cel puțin ar trebui să fie uniformă în întregul grup.

3.2.2.3. Limitele l-diversității

În tabelul de mai jos, l-diversitatea este aplicată în ceea ce privește atributul „Diagnostic"; cu toate acestea, știind că o persoană născută în 1964 figurează în acest tabel, este în continuare posibil să se presupună cu o probabilitate foarte mare că aceasta a suferit un atac de cord.

Anul	Sex	Codul poștal	Diagnostic
1957	M	750*	**Atac de cord**
1957	M	750*	**Colesterol**
1957	M	750*	**Colesterol**
1957	M	750*	**Colesterol**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Colesterol**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**
1964	M	750*	**Atac de cord**

Tabelul 3. Un tabel l-divers în care valorile pentru „Diagnostic" nu sunt distribuite uniform

Nume	Data nașterii	Sex
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Tabelul 4. Știind că aceste persoane figurează în tabelul 3, un atacator ar putea deduce că acestea au suferit un atac de cord

4. Pseudonimizarea

Pseudonimizarea constă în înlocuirea unui atribut (de regulă, un atribut unic) în cadrul unei înregistrări cu un altul. Prin urmare, persoana fizică încă poate fi identificată în mod indirect; în consecință, pseudonimizarea, atunci când este utilizată separat, nu va avea drept rezultat un set de date anonim. Cu toate acestea, pseudonimizarea este abordată în prezentul aviz ca urmare a numeroaselor concepții greșite și erori legate de utilizarea acesteia.

Pseudonimizarea reduce posibilitatea de a crea legături în cadrul unui set de date cu identitatea originală a unei persoane vizate; ca atare, aceasta reprezintă o măsură de securitate utilă, însă nu și o metodă de anonimizare.

Rezultatul pseudonimizării poate fi independent de valoarea inițială (astfel cum este cazul unui număr aleatoriu generat de operator sau al unui nume ales de persoana vizată) sau poate fi derivat din valorile inițiale ale unui atribut sau ale unui set de atribute de exemplu, o funcție hash sau un sistem de criptare.

Cele mai utilizate tehnici de pseudonimizare sunt după cum urmează:

- criptarea cu o cheie secretă: în acest caz, deținătorul cheii poate cu ușurință să reidentifice fiecare persoană vizată prin decriptarea setului de date deoarece datele cu caracter personal încă fac parte din setul de date, chiar dacă sub o formă criptată. Presupunând că s-a aplicat un sistem de criptare de ultimă generație, decriptarea poate fi posibilă numai dacă se cunoaște cheia.
- funcția *hash*: aceasta corespunde unei funcții care furnizează un rezultat cu o dimensiune fixă dintr-o intrare de orice dimensiuni (intrarea poate fi formată dintr-un singur atribut sau un set de atribute) și nu poate fi inversată; acest lucru înseamnă că riscul de inversare asociat criptării nu mai există. Cu toate acestea, dacă intervalul de valori de intrare ale funcției hash sunt cunoscute, acestea pot fi repetate prin intermediul funcției hash pentru a obține valoarea corectă a unei anumite înregistrări. De exemplu, dacă un set de date a fost pseudonimizat prin hashing-ul numărului național de identificare, atunci acesta poate fi obținut aplicând hashing-ul tuturor valorilor de intrare și apoi comparând rezultatul cu valorile respective din setul de date. Funcțiile hash sunt concepute, de regulă, să fie relativ ușor de calculat și fac

obiectul unor atacuri prin forță brută[16]. De asemenea, pot fi create tabele calculate în prealabil pentru a permite inversarea în masă a unui set amplu de valori hash.

Utilizarea unei funcții de tip *salted-hash* (în care o valoare aleatorie, cunoscută sub denumirea de „hash inteligent", se adaugă la atributul căruia urmează să i se aplice hashing-ul) poate reduce probabilitatea de a se obține valoarea inițială însă, cu toate acestea, calcularea valorii atributului inițial ascunsă în spatele rezultatului unei funcții de tip *salted hash* poate fi în continuare realizabilă cu ajutorul unor mijloace rezonabile[17].

- funcție de tip *keyed-hash* cu o cheie stocată: aceasta corespunde unei funcții hash speciale care folosește o cheie secretă ca element de intrare suplimentar (aceasta diferă de funcția de tip *salted hash,* deoarece, de regulă, valoarea *salt* nu este secretă). Un operator de date poate aplica din nou funcția pe atribut utilizând cheia secretă, însă este mult mai dificil pentru un atacator să repete funcția fără să cunoască cheia, întrucât numărul de posibilități care trebuie testate este suficient de mare pentru a face acest lucru imposibil de aplicat.

- criptarea deterministă sau funcția de tip *keyed-hash* cu ștergerea cheii: această tehnică poate fi echivalată cu selectarea unui număr aleatoriu drept pseudonim pentru fiecare atribut din baza de date și apoi cu ștergerea tabelului de corespondență. Această soluție permite[18] diminuarea riscului de stabilire de legături între datele cu caracter personal din setul de date și cele referitoare la aceeași persoană dintr-un alt set de date unde este folosit un pseudonim diferit. Având în vedere un algoritm de ultimă generație, va fi dificil din punct de vedere al realizării calculelor pentru un atacator să decripteze sau să repete funcția, întrucât acest lucru ar presupune testarea fiecărei chei posibile, având în vedere că respectiva cheie nu este disponibilă.

- tokenizare: această tehnică este aplicată, de regulă, (chiar dacă nu se limitează la acesta) în sectorul financiar pentru înlocuirea numerelor de identificare ale cardurilor cu valori care au o utilitate redusă pentru un atacator. Tehnica rezultă din cele precedente, fiind bazată, de regulă, pe aplicarea unor mecanisme de criptare unidirecțională sau atribuirea, prin intermediul unui funcții de index, unui număr secvențial sau a unui număr generat în mod aleatoriu, care nu este obținut în mod matematic din datele originale.

4.1. Garanții

- *Individualizarea:* există în continuare posibilitatea de a se individualiza înregistrările cu privire la o persoană, întrucât aceasta este identificată în continuare printr-un atribut unic care este rezultatul funcției de pseudonimizare (= atribut pseudonimizat).

- *Posibilitatea stabilirii de legături:* stabilirea de legături va fi în continuare ușoară între înregistrările care utilizează același atribut pseudonimizat pentru a face referire la aceeași persoană. Chiar dacă sunt utilizate atribute pseudonimizate diferite pentru aceeași persoană vizată, stabilirea de legături poate fi posibilă în continuare prin intermediul altor atribute. Nu va exista o referință încrucișată evidentă între două seturi de date care folosesc diferite atribute pseudonimizate doar în cazul în care

[16] Astfel de atacuri constau în încercarea tuturor intrărilor posibile în vederea elaborării de tabele de corespondență.
[17] În special, dacă se cunoaște tipul de atribut (nume, număr social, data nașterii etc.). Pentru adăugarea unei cerințe de calcul, se poate lua drept bază o funcție hash de derivare a cheii în care valoarea calculată este supusă de mai multe ori unui hashing cu o valoare *salt* mică.
[18] În funcție de celelalte atribute din setul de date și de ștergerea datelor inițiale.

niciun alt atribut din setul de date nu poate fi utilizat pentru a se identifica persoana vizată și dacă fiecare legătură dintre atributul original și atributul pseudonimizat a fost eliminată (inclusiv prin ștergerea datelor originale).

- *Deducția:* atacurile prin deducție cu privire la identitatea reală a unei persoane vizate sunt posibile în cadrul setului de date sau în rândul diferitelor seturi de date care folosesc același atribut pseudonimizat pentru o persoană sau dacă pseudonimele sunt explicite și nu maschează în mod corespunzător identitatea originală a persoanei vizate.

4.2. Erori frecvente

- Convingerea că un set de date pseudonimizate este anonimizat: adeseori operatorii de date presupun că eliminarea sau înlocuirea unuia sau a mai multor atribute este suficientă pentru a face ca setul de date să devină anonim. Numeroase exemple au arătat că acest lucru nu este valabil; simpla modificare a identității nu împiedică o persoană să identifice o persoană vizată în cazul în care cvasi-identificatorii rămân în setul de date sau dacă valorile altor atribute permit în continuare identificarea unei persoane. În multe cazuri, poate fi la fel de ușor să se identifice o persoană într-un set de date pseudonimizate precum în cazul datelor originale. Ar trebui luate măsuri suplimentare pentru a putea considera un set de date ca fiind anonimizat, inclusiv eliminarea și generalizarea atributelor sau ștergerea datelor originale sau cel puțin aducerea acestora la un nivel ridicat de agregare.
- Erori frecvente atunci când se utilizează pseudonimizarea ca tehnică pentru a se reduce posibilitatea stabilirii de legături:
    - utilizarea aceleiași chei în baze de date diferite: eliminarea posibilității de a se stabili legături în cadrul diferitelor seturi de date depinde în mare măsură de utilizarea unui algoritm cu cheie și de faptul că o singură persoană va corespunde unor atribute pseudonimizate diferite în contexte diferite. Prin urmare, este important să se evite utilizarea aceleiași chei în diferite baze de date pentru a putea reduce posibilitatea creării de legături.
    - utilizarea unor chei diferite („chei rotative”) pentru utilizatori diferiți: ar putea fi tentant să se utilizeze chei diferite pentru seturi diferite de utilizatori și să se modifice cheia la fiecare utilizare (de exemplu, utilizarea aceleiași chei pentru a înregistra 10 intrări referitoare la același utilizator). Cu toate acestea, dacă nu este elaborată corect, această operațiune ar putea determina apariția unor modele, reducând parțial beneficiile preconizate. De exemplu, rotirea cheii prin norme specifice pentru anumite persoane ar facilita posibilitatea creării de legături între intrările corespunzătoare anumitor persoane. De asemenea, dispariția unor date pseudonimizate recurente din baza de date în momentul apariției altora noi ar putea indica faptul că ambele înregistrări se referă la aceeași persoană fizică.
    - păstrarea cheilor: în cazul în care cheia secretă este păstrată împreună cu datele pseudonimizate, iar datele sunt compromise, atacatorul poate asocia cu ușurință datele pseudonimizate cu atributul lor original. Același lucru este valabil în cazul în care cheia este stocată separat de date, dar nu în mod securizat.

4.3. Punctele slabe ale pseudonimizării

- *Asistența medicală*

1. Nume, adresă, data nașterii	2. Perioada de beneficiere de asistență specială	3. Indicele de masă corporală	6. Numărul de referință al echipei de cercetare
[redacted]	< 2 ani	15	QA5FRD4
	> 5 ani	14	2B48HFG
	< 2 ani	16	RC3URPQ
	> 5 ani	18	SD289K9
	< 2 ani	20	5E1FL7Q

Tabelul 5. Un exemplu de pseudonimizare prin hashing (nume, adresă, data nașterii) care poate fi inversată cu ușurință

Un set de date a fost creat pentru a examina relația dintre greutatea unei persoane și plata unui beneficiu de asistență specială. Setul de date original a inclus numele, adresa și data nașterii persoanelor vizate, însă acestea au fost șterse. Numărul de referință al echipei de cercetare a fost generat din datele șterse folosindu-se o funcție hash. Deși numele, adresa și data nașterii au fost șterse din tabel, în cazul în care numele, adresa sau data nașterii unei persoane sunt cunoscute, pe lângă cunoașterea funcției hash utilizate, este ușor să se calculeze numerele de referință ale echipelor de cercetare.

- *Rețele sociale*

S-a demonstrat[19] că informații sensibile cu privire la anumite persoane pot fi extrase din grafice de tip rețea socială, în pofida tehnicilor de pseudonimizare aplicate datelor respective. Un furnizor de rețea socială a presupus în mod eronat că pseudonimizarea a fost solidă pentru a preveni identificarea după vânzarea datelor către alte societăți în scopuri de marketing și de publicitate. În locul numelor reale, furnizorul a utilizat porecle însă, în mod evident, acest lucru nu a fost suficient pentru a anonimiza profilurile de utilizator deoarece relațiile dintre diferitele persoane sunt unice și pot fi utilizate ca element de identificare.

- *Locații*

Cercetătorii de la MIT[20] au analizat recent un set de date pseudonimizate constând în coordonate de mobilitate spațial-temporală pe o perioadă de 15 luni pentru un număr de 1,5 milioane de oameni pe un teritoriu cu o rază de 100 km. Aceștia au demonstrat că 95 % din populație putea fi individualizată utilizând-se patru puncte de localizare și că doar două puncte erau suficiente pentru a se individualiza peste 50 % din persoanele vizate (unul dintre aceste puncte este cunoscut, fiind foarte probabil „acasă” sau „serviciu”), lăsându-se un spațiu foarte limitat pentru protecția vieții private, chiar dacă identitățile persoanelor au fost pseudonimizate prin înlocuirea atributelor lor reale [….] cu alte etichete.

[19] A. Narayanan și V. Shmatikov, „De-anonymizing social networks”, 30th IEEE Symposium on Security and Privacy, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen și V. Blondel, „Unique in the Crowd: The privacy bounds of human mobility”, Nature, no. 1376, 2013.

5. Concluzii și recomandări

5.1. Concluzii

Tehnicile de eliminare a posibilității de identificare și de anonimizare fac obiectul unor cercetări intense, iar prezentul document a arătat în mod consecvent că fiecare tehnică are avantajele și dezavantajele sale. În cele mai multe cazuri, nu este posibil să se furnizeze recomandări minime cu privire la parametrii care urmează să fie utilizați deoarece fiecare set de date trebuie să fie analizat de la caz la caz.

În multe cazuri, un set de date anonimizat poate prezenta în continuare un risc rezidual pentru persoanele vizate. Într-adevăr, inclusiv în cazul în care nu mai este posibilă recuperarea cu exactitate a înregistrării cu privire la o persoană, poate rămâne posibilă spicuirea de informații cu privire la o persoană cu ajutorul altor surse de informații care sunt disponibile (în mod public sau nu). Trebuie subliniat faptul că, pe lângă impactul direct asupra persoanelor vizate produs de consecințele unui proces de anonimizare deficitar (iritare, timp pierdut și sentimentul de pierdere a controlului prin includerea într-un grup fără a fi informat sau fără un consimțământ prealabil), pot apărea și alte efecte secundare ale anonimizării deficitare atunci când persoana vizată este inclusă într-o țintă, în mod eronat, de către un atacator, ca urmare a prelucrării datelor anonimizate – mai ales dacă intențiile atacatorului sunt răuvoitoare. Prin urmare, grupul de lucru subliniază că tehnicile de anonimizare pot oferi garanții privind viața privată, dar numai dacă aplicarea lor este concepută în mod corespunzător – ceea ce înseamnă că trebuie să fie stabilite în mod clar condițiile prealabile (contextul) și obiectivul (obiectivele) procesului de anonimizare pentru a se atinge nivelul vizat de anonimizare.

5.2. Recomandări

- Unele tehnici de anonimizare prezintă limitări inerente. Limitările respective trebuie să fie luate în considerare cu seriozitate de către operatorii de date înainte de a utiliza o anumită tehnică de anonimizare pentru a elabora un proces de anonimizare. Aceștia trebuie să aibă în vedere scopurile care trebuie realizate prin anonimizare – cum ar fi protecția vieții private a persoanelor atunci când publică un set de date sau când permit recuperarea unei informații dintr-un set de date.
- Niciuna dintre tehnicile descrise în prezentul document nu îndeplinește cu certitudine criteriile unei anonimizări eficiente (și anume, imposibilitatea identificării unei persoane; imposibilitatea stabilirii de legături între înregistrările referitoare la o persoană și imposibilitatea deducției cu privire la o persoană). Cu toate acestea, întrucât unele dintre aceste riscuri pot fi abordate în întregime sau parțial de către o anumită tehnică, este necesară o inginerie atentă în conceperea aplicării unei tehnici individuale într-o situație specifică și în aplicarea unei combinații a acestor tehnici ca o modalitate de a consolida soliditatea rezultatului.

Tabelul de mai jos oferă o prezentare generală a punctelor tari și a punctelor slabe ale tehnicilor analizate în ceea ce privește cele trei cerințe de bază:

	Există în continuare riscul identificării?	Există în continuare riscul de a se stabili legături?	Există în continuare riscul deducției?
Pseudonimizare	Da	Da	Da
Adăugarea de zgomot	Da	Posibil nu	Posibil nu
Substituție	Da	Da	Posibil nu
Agregare sau k-anonimat	Nu	Da	Da
L-diversitate	Nu	Da	Posibil nu
Confidențialitate diferențială	Posibil nu	Posibil nu	Posibil nu
Hashing/tokenizare	Da	Da	Posibil nu

Tabelul 6. Punctele tari și punctele slabe ale tehnicilor analizate

- Soluția optimă ar trebui să fie decisă de la caz la caz. O soluție (și anume, un proces complet de anonimizare) care îndeplinește cele trei criterii ar fi o soluție solidă împotriva identificării efectuate prin cele mai probabile și rezonabile mijloace pe care operatorul de date sau oricare altă parte terță le poate utiliza.
- Ori de câte ori o propunere nu îndeplinește unul dintre criterii, ar trebui efectuată o evaluare aprofundată a riscurilor de identificare. Evaluarea ar trebui să fie transmisă autorității în cazul în care legislația națională prevede faptul că autoritatea trebuie să evalueze sau să autorizeze procesul de anonimizare.

Pentru a reduce riscurile de identificare, următoarele bune practici ar trebui să fie luate în considerare:

<u>Bune practici privind anonimizarea</u>

În general:

- Operatorii de date nu ar trebui să se bazeze pe abordarea „publică și uită”. Având în vedere riscul rezidual de identificare, aceștia ar trebui:
    - 1. să identifice noile riscuri și să reevalueze riscul (riscurile) rezidual(e) în mod regulat,
    - 2. să evalueze dacă verificările efectuate pentru riscurile identificate sunt suficiente și să le adapteze în consecință; ȘI
    - 3. să monitorizeze și să controleze riscurile.
- Ca parte a unor astfel de riscuri reziduale, operatorii de date ar trebui să ia în considerare potențialul de identificare a segmentului non-anonimizat al unui set de date (în cazul în care există), în special atunci când este combinat cu segmentul anonimizat, precum și posibilele corelații între atribute (de exemplu, între localizarea geografică și datele privind nivelul de bunăstare).

Elemente contextuale:

- Scopurile care trebuie atinse prin intermediul setului de date anonimizate ar trebui să fie stabilite în mod clar deoarece acestea joacă un rol esențial în determinarea riscului de identificare.

- Acest aspect este strâns legat de examinarea tuturor elementelor contextuale relevante – de exemplu, natura datelor originale, mecanisme de control instituite (inclusiv măsurile de securitate pentru a limita accesul la seturile de date), dimensiunea eșantionului (caracteristici cantitative), disponibilitatea resurselor de informații publice (pe care să se poată baza destinatarii), publicarea preconizată de date către părți terțe (limitată, nelimitată, de exemplu pe internet etc.).
- Ar trebui avuți în vedere posibilii atacatori prin luarea în considerare a atractivității datelor pentru atacurile specifice (din nou, sensibilitatea informațiilor și natura datelor vor fi factori esențiali în acest sens).

Elemente tehnice:

- Operatorii de date ar trebui să dezvăluie tehnica de anonimizare/combinația de tehnici aplicată, în special în cazul în care intenționează să publice setul de date anonimizat.
- Ar trebui să se elimine din setul de date atributele (de exemplu, cele rare) evidente /cvasi-identificatorii.
- În cazul în care se folosesc tehnici de adăugare de zgomot (în randomizare), nivelul de zgomot adăugat înregistrărilor ar trebui să fie stabilit în funcție de valoarea unui atribut (și anume, nu ar trebui să se injecteze un zgomot disproporționat), impactul pentru persoanele vizate al atributelor care urmează să fie protejate și/sau raritatea setului de date.
- Atunci când se recurge la confidențialitatea diferențială (în randomizare), operatorii de date ar trebui să se țină seama de necesitatea de a ține evidența interogărilor astfel încât să se detecteze interogările care constituie o ingerință în viața privată, întrucât ingerința interogărilor este cumulativă.
- În cazul în care sunt aplicate tehnici de generalizare, este fundamental pentru operatorul de date să nu se limiteze la un singur criteriu de generalizare chiar și pentru același atribut; cu alte cuvinte, ar trebui selectate puncte de localizare diferite sau intervale de timp diferite. Alegerea criteriului care urmează să fie aplicat trebuie să fie determinată de distribuția valorilor atributelor în populația respectivă. Nu toate distribuțiile sunt susceptibile de generalizare – ceea ce înseamnă că nu există o abordare general valabilă care să poată fi aplicată în generalizare. Ar trebui asigurată variabilitatea în cadrul claselor de echivalență; de exemplu, ar trebui selectat un anumit prag în funcție de „elementele contextuale” menționate mai sus (dimensiunea eșantionului etc.) și, în cazul în care acesta nu este atins, ar trebui eliminat eșantionul specific (sau ar trebui stabilit un alt criteriu de generalizare).

ANEXĂ

Prezentare generală a tehnicilor de anonimizare

A. 1. Introducere

Anonimatul este interpretat în mod diferit în UE – în unele țări acesta corespunde anonimatului de calcul (și anume, ar trebui să fie dificil din punct de vedere al realizării calculelor, chiar și pentru operator în colaborare cu oricare parte terță, să identifice, în mod direct sau indirect, una dintre persoanele vizate), iar în alte țări acesta corespunde anonimatului perfect (și anume, ar trebui să fie imposibil, chiar și pentru operator în colaborare cu oricare parte terță, să identifice, în mod direct sau indirect, una dintre persoanele vizate). Cu toate acestea, „anonimizarea" corespunde în ambele cazuri procesului prin care datele devin anonime. Diferența constă în ceea ce se consideră a fi un nivel acceptabil al riscului de reidentificare a persoanei vizate.

În ceea ce privește datele anonimizate, pot fi avute în vedere diferite cazuri de utilizare, variind de la anchete sociale, analize statistice, la dezvoltarea de noi servicii/produse. Uneori chiar și astfel de activități cu un obiectiv general pot avea un impact asupra unor persoane vizate specifice, ceea ce anulează presupusa natură anonimă a datelor prelucrate. În acest sens pot fi menționate numeroase exemple, de la lansarea unor inițiative de marketing orientate, până la punerea în aplicare a unor măsuri publice bazate pe crearea de profiluri de utilizator sau pe comportamente sau modele de mobilitate[21].

Din păcate, dincolo de afirmațiile generale, nu există unități de măsură dezvoltate pentru a se putea evalua în prealabil timpul sau efortul necesare pentru reidentificare după etapa de prelucrare sau, în mod alternativ, pentru a se alege procedura cea mai adecvată de punere în aplicare în cazul în care se dorește reducerea probabilității ca baza de date publicată să se refere la un set identificat de persoane vizate.

„Arta anonimizării", astfel cum sunt uneori denumite aceste practici în literatura de specialitate[22], constituie o nouă ramură științifică aflată încă în perioada de început, existând numeroase practici pentru diminuarea puterii de identificare a seturilor de date; cu toate acestea, trebuie să se menționeze în mod clar că majoritatea acestor practici nu împiedică stabilirea de legături între datele prelucrate și persoanele vizate. În anumite circumstanțe, identificarea seturilor de date considerate anonime s-a dovedit a avea un grad foarte ridicat de succes, în timp ce în alte circumstanțe aceasta s-a dovedit fals pozitivă.

În general, există două abordări diferite: una se bazează pe generalizarea atributelor, cealaltă pe randomizare. Analiza detaliilor și a subtilităților acestor practici ne va oferi o nouă perspectivă asupra puterii de identificare a datelor și va arunca o nouă lumină asupra noțiunii însăși de date cu caracter personal.

A.2. „Anonimizarea" prin randomizare

O opțiune pentru anonimizare constă în modificarea valorilor reale pentru a se preveni stabilirea de legături între datele anonimizate și valorile originale. Acest obiectiv poate fi atins prin intermediul mai multor metodologii, variind de la injectarea de zgomot la schimbul de date (permutare). Trebuie subliniat faptul că eliminarea unui atribut este echivalentă cu o formă extremă de randomizare a atributului respectiv (atributul fiind acoperit în întregime de zgomot).

[21] De exemplu, cazul TomTom în Țările de Jos (a se vedea exemplul explicat la punctul 2.2.3).
[22] Jun Gu,Yuexian Chen, Junning Fu, HuanchunPeng, Xiaojun Ye, Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes in Computer Science –Springer – Volume 6261, 2010, p. 385-399.

În unele cazuri, obiectivul prelucrării globale nu este neapărat emiterea unui set de date randomizate, ci mai degrabă acordarea accesului la date prin intermediul interogărilor. În acest caz, riscul pentru persoana vizată rezultă din probabilitatea ca un atacator să poată extrage informații printr-o serie de interogări diverse, fără ca operatorul de date să aibă cunoștință de acest lucru. Pentru a se garanta anonimatul persoanelor din setul de date cu privire la acestea, nu ar trebui să fie posibil să se poată concluziona că o persoană vizată a contribuit la setul de date, rupând astfel legătura cu orice fel de informații generale pe care un atacator le-ar putea deține.

Adăugarea de zgomot, după caz, la răspunsul obținut în urma interogării, poate reduce și mai mult riscul de reidentificare. Această abordare, cunoscută de asemenea, în literatura de specialitate sub denumirea de confidențialitate diferențială[23], are ca punct de plecare cele descrise anterior în sensul că oferă editorilor de date un grad mai ridicat de control asupra accesului la date comparativ cu publicarea acestora. Adăugarea de zgomot are două obiective principale: primul, de a proteja viața privată a persoanelor vizate din setul de date și al doilea, de a păstra utilitatea informațiilor publicate. În special, amploarea zgomotului trebuie să fie proporțională cu nivelul de interogare (prea multe întrebări cu privire la persoane la care să se răspundă cu prea multă precizie au drept rezultat creșterea probabilității de identificare). În prezent, aplicarea cu succes a randomizării trebuie să fie analizată de la caz la caz, întrucât nicio tehnică nu oferă o metodologie infailibilă; există exemple de scurgeri de informații privind atributele unei persoane vizate (incluse sau nu în setul de date), inclusiv atunci când setul de date a fost considerat randomizat de către operatorul de date.

Ar putea fi util să se discute exemple concrete pentru a clarifica potențiale limite ale randomizării ca mijloc de a garanta anonimizarea. De exemplu, în contextul accesului interactiv, interogările considerate ca respectând viața privată ar putea prezenta un risc pentru persoanele vizate. De fapt, dacă atacatorul cunoaște că un subgrup S de persoane se află într-un set de date care conține informații privind incidența atributului A într-o populație P, prin simpla interogare cu două întrebări „Câte persoane din populația P dețin atributul A?” și „Câte persoane din populația P, cu excepția celor aparținând subgrupului S, dețin atributul A?”, se poate determina (prin diferență) numărul de persoane din S care dețin de fapt atributul A – fie în mod determinist, fie prin deducția prin probabilitate. În orice caz, viața privată a persoanelor din subgrupul S ar putea fi pusă serios în pericol, în special, în funcție de natura atributului A.

De asemenea, se poate considera că, dacă o persoană vizată nu este în setul de date, dar se cunoaște legătura acestuia cu datele din setul de date, atunci publicarea setului de date poate cauza un risc pentru viața privată a persoanei. De exemplu, dacă se cunoaște faptul că „valoarea atributului A al țintei diferă cu o cantitate X de valoarea medie a populației”, prin simpla solicitare a custodelui bazei de date de a efectua o operațiune care respectă dreptul la viață privată de a extrage valoarea medie a atributului A, atacatorul poate deduce cu exactitate date cu caracter personal referitoare la o anumită persoană vizată.

Injectarea unor inexactități relative printre valorile reale dintr-o bază de date constituie o operațiune care ar trebui să fie concepută în mod corespunzător. Este necesar să se adauge suficient de multe elemente de zgomot pentru a se proteja viața privată, însă și suficient de puține pentru a se păstra utilitatea datelor. De exemplu, în cazul în care numărul de persoane vizate cu un atribut deosebit este foarte mic sau sensibilitatea atributului este mare, ar putea fi mai bine să se raporteze o serie sau o afirmație generică precum „un număr mic de cazuri,

[23] Cynthia Dwork, Differential Privacy, International Colloquium on Automata, Languages and Programming (ICALP) 2006, p. 1-12.

chiar zero", în loc de raportarea numărului real. În acest mod, chiar dacă se cunoaște dinainte mecanismul zgomotos de divulgare, viața privată a persoanei vizate este protejată, întrucât rămâne un grad de incertitudine. Dintr-o perspectivă a utilității, în cazul în care inexactitatea este concepută în mod corespunzător, rezultatele sunt în continuare utile pentru scopuri statistice sau de luare a deciziilor.

Randomizare bazei de date și accesul la confidențialitatea diferențială necesită o reflecție suplimentară. În primul rând, cantitatea corectă de denaturare poate varia în mod semnificativ în funcție de context (tip de interogare, dimensiunea populației din baza de date, natura atributului și puterea sa inerentă de identificare) și nu poate fi avută în vedere o soluție „ad omnia". De asemenea, contextul se poate schimba în timp, iar mecanismul interactiv ar trebui modificat în consecință. Calibrarea zgomotului necesită monitorizarea riscurilor cumulative pentru dreptul la viață privată pe care orice mecanism interactiv le implică pentru persoanele vizate. Prin urmare, mecanismul de acces la date ar trebui să fie echipat cu sisteme de alertă atunci când a fost atins bugetul pentru „costul vieții private" și persoanele vizate ar putea fi expuse unor riscuri specifice în cazul în care este formulată o nouă interogare, pentru a sprijini operatorul de date să determine nivelului adecvat de denaturare pe care trebuie să-l injecteze de fiecare dată în datele reale cu caracter personal.

Pe de altă parte, ar trebui să se aibă în vedere, de asemenea, cazul în care valorile atributelor sunt șterse (sau modificate). O soluție utilizată în mod curent pentru a aborda unele valori atipice ale atributelor este ștergerea fie a setului de date referitoare la persoane atipice, fie a valorilor atipice. În cel din urmă caz, este important să se asigure că absența valorii în sine nu devine un element de identificare a unei persoane vizate.

Să analizăm în continuare randomizarea prin substituția atributului. O concepție greșită importantă atunci când se are în vedere anonimizarea este aceea de a o echivala cu criptarea sau cu codarea cu cheie. Această concepție greșită se bazează pe două ipoteze, și anume, a) că din momentul în care s-a aplicat criptarea anumitor atribute ale unei înregistrări dintr-o bază de date (de exemplu, nume, adresă, data nașterii) sau aceste atribute sunt înlocuite cu un șir aparent randomizat ca urmare a unei operațiuni de codare cu cheie precum funcția de tip keyed-hash, înregistrarea respectivă este „anonimizată" și b) că anonimizarea este mai eficientă dacă lungimea cheii este adecvată, iar algoritmul de criptare este de ultimă generație. Această concepție greșită este răspândită în rândul operatorilor de date și merită clarificări deoarece ea este legată, de asemenea, de pseudonimizare și presupusele sale riscuri mai mici.

În primul rând, obiectivele acestor tehnici sunt radical diferite: criptarea ca practică de securitate urmărește să asigure confidențialitatea unui canal de comunicare între părți identificate (persoane, dispozitive sau piese de software și hardware) pentru a evita interceptarea sau divulgarea neintenționată. Codarea cu cheie corespunde unei traduceri semantice a datelor în funcție de o cheie secretă. Dimpotrivă, obiectivul anonimizării este de a evita identificarea persoanelor prin prevenirea stabilirii de legături ascunse între atribute și o persoană vizată.

Nici criptarea, nici codarea cu cheie nu se pretează în sine la obiectivul de a face ca o persoană vizată să devină non-identificabilă: ca atare, cel puțin în ceea ce privește operatorul, datele originale sunt în continuare disponibile sau deductibile. Simpla implementare a unei traduceri semantice a datelor cu caracter personal, astfel cum se întâmplă în cazul codării cu cheie, nu elimină posibilitatea de a restabili structura inițială a datelor — fie prin aplicarea algoritmului în sens invers, fie prin atacuri prin forță brută, în funcție de natura sistemelor în cauză sau ca urmare a unei încălcări a securității datelor. Criptarea de ultimă generație poate garanta că datele sunt protejate într-o măsură mai mare, și anume, acestea sunt neinteligibile

pentru entitățile care nu cunosc cheia de decriptare, însă nu are în mod necesar drept rezultat anonimizarea. Atât timp cât cheia sau datele originale sunt disponibile (inclusiv în cazul unei părți terțe de încredere, obligată prin contract să furnizeze servicii cheie sigure), posibilitatea de a identifica o persoană vizată nu este eliminată.

Axarea doar pe soliditatea mecanismului de criptare ca măsură a gradului de „anonimizare" a unui set de date este înșelătoare, întrucât numeroși alți factori tehnici și organizatorici afectează securitatea globală a unui mecanism de criptare sau a unei funcții hash. În literatura de specialitate au fost semnalate numeroase atacuri reușite care evită în totalitate algoritmul, fie datorită faptului că acestea profită de pe urma deficiențelor în custodia cheilor (de exemplu, existența unui mod implicit mai puțin sigur), fie datorită altor factori umani (de exemplu, parole slabe pentru recuperarea cheii). În sfârșit, un sistem de criptare ales cu o anumită dimensiune a cheii este menit să asigure confidențialitatea pentru o anumită perioadă (majoritatea cheilor actuale vor trebui să fie redimensionate în jurul anului 2020), în timp ce un proces de anonimizare nu ar trebui să fie limitat în timp.

În continuare, ar trebui analizate limitele randomizării atributelor (sau substituția și eliminarea), luându-se în considerare diferitele exemple negative de anonimizare prin randomizare care au avut loc în ultimii ani și motivele care stau la baza eșecului acestora.

Un caz binecunoscut care implică publicarea unui set de date anonimizate ineficient este cel al Premiului Netflix[24]. Analizând o înregistrare generică dintr-o bază de date în care un număr de atribute referitoare la o persoană vizată au fost randomizate, fiecare înregistrare poate fi în continuare divizată în două sub-înregistrări, după cum urmează: {atribute randomizate, atribute clare}, unde atributele clare pot reprezenta orice combinație de presupuse date fără caracter personal. O observație specifică ce se poate efectua pe baza setului de date privind Premiul Netflix este generată de constatarea că fiecare înregistrare poate fi reprezentată de un punct într-un spațiu multidimensional, unde fiecare atribut clar reprezintă o coordonată. Prin utilizarea acestei tehnici, orice set de date poate fi considerat ca o constelație de puncte într-un astfel de spațiu multidimensional care poate prezenta un grad ridicat de raritate, însemnând că punctele sunt/se pot afla la distanță unele de celelalte. Într-adevăr, acestea pot fi atât de îndepărtate încât, după divizarea spațiului în regiuni mari, fiecare regiune poate să conțină numai o înregistrare. Chiar și injectarea de zgomot nu reușește să aducă înregistrările suficient de aproape încât să facă parte din aceeași regiune multidimensională. De exemplu, în experimentul Netflix, înregistrările au fost suficient de singulare, cu doar 8 evaluări de filme acordate la o distanță de 14 zile. După adăugarea de zgomot atât în cazul evaluărilor, cât și în cazul datelor nu s-a putut observa o suprapunere a regiunilor. Cu alte cuvinte, aceeași selecție de numai 8 filme evaluate a constituit amprenta evaluărilor acordate, nefiind partajate de două persoane vizate din baza de date. Pe baza acestei observații geometrice, cercetătorii au asociat setul de date Netflix presupus anonim cu o altă bază de date publică cu evaluări de filme (IMDB), identificând astfel utilizatori care au acordat evaluările pentru aceleași filme în aceleași intervale de timp. Având în vedere că majoritatea utilizatorilor au prezentat o corespondență univocă, informațiile auxiliare extrase din baza de date IMDB au putut fi importate în setul de date Netflix publicat, îmbogățind astfel cu identități toate presupusele înregistrări anonime.

Este important de subliniat faptul că aceasta este o proprietate generală: partea reziduală a oricărei baze de date „randomizate" prezintă în continuare o putere foarte ridicată de identificare, în funcție de raritatea combinației atributelor reziduale. Acesta este un neajuns pe

[24] Arvind Narayanan, Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, IEEE Symposium on Security and Privacy 2008: 111-125.

care operatorii de date ar trebui să-l aibă mereu în vedere atunci când selectează randomizarea ca modalitate prin care intenționează să realizeze anonimizarea.

Numeroase experimente de reidentificare de acest tip au urmat o abordare similară cu cea a proiectării a două baze de date pe același subspațiu. Acesta este o metodologie solidă de reidentificare care a avut recent numeroase aplicații în domenii diferite. De exemplu, un experiment de identificare efectuat în raport cu o rețea socială[25] a exploatat graficul social al utilizatorilor pseudonimizați cu ajutorul etichetelor. În acest caz, atributele utilizate pentru identificare au fost lista contactelor fiecărui utilizator, întrucât s-a demonstrat că probabilitatea existenței unei liste identice a contactelor în cazul a două persoane este foarte scăzută. Pe baza acestei ipoteze intuitive, s-a constatat că un sub-grafic al conexiunilor interne ale unui număr foarte limitat de noduri constituie o amprentă topologică care poate fi extrasă, ascunsă în cadrul rețelei, și că o mare parte din întreaga rețea socială poate fi identificată ulterior identificării acestei sub-rețele. Doar în scopul de a se furniza câteva cifre cu privire la performanțele unui atac similar, s-a arătat că prin utilizarea a mai puțin de 10 noduri (ceea ce poate da naștere la un milion de configurații diferite de sub-rețele, fiecare dintre acestea reprezentând o potențială amprentă topologică) o rețea socială de peste 4 milioane de noduri pseudonimizate și 70 de milioane de legături poate fi vulnerabilă la atacuri privind reidentificarea, iar confidențialitatea unui număr mare de legături poate fi compromisă. Trebuie subliniat că această abordare a reidentificării nu este adaptată contextului specific al rețelelor sociale, însă este suficient de generală pentru a putea fi eventual adaptată la alte baze de date unde relațiile dintre utilizatori sunt înregistrate (de exemplu, contacte telefonice, corespondența electronică, site-uri de matrimoniale etc.).

O altă modalitate de a identifica o presupusă înregistrare anonimă se bazează pe analiza stilului de redactare (stilometrie)[26]. O serie de algoritmi au fost deja dezvoltați pentru a extrage date metrice dintr-un text analizat inclusiv frecvența utilizării unui anumit cuvânt, apariția unor modele gramaticale specifice și tipul de punctuație. Toate aceste proprietăți pot fi utilizate pentru a ancora un presupus text anonimizat în stilul de redactare al unui autor identificat. Cercetătorii au extras stilul de redactare al unui număr de peste 100 000 de bloguri și, în prezent, sunt în măsură să identifice în mod automat autorul unei postări cu o precizie care se apropie deja de 80 %; se preconizează că precizia acestei tehnici va crește în continuare prin exploatarea, de asemenea, a altor indicii precum locația sau alte metadate conținute în text.

Puterea de identificare prin utilizarea semanticii unei înregistrări (și anume, partea reziduală nerandomizată a unei înregistrări) constituie un aspect care necesită o mai mare atenție din partea comunității de cercetare și din partea industriei. Recenta inversare a identităților donatorilor de ADN (2013)[27] arată că s-au înregistrat foarte puține progrese de la binecunoscutul incident AOL (2006) – când o bază de date cuprinzând douăzeci de milioane de cuvinte cheie pentru peste 650 000 de utilizatori pe o perioadă de 3 luni a fost făcută publică. Aceasta a condus la identificarea și localizarea unui număr de utilizatori AOL.

[25] L. Backstrom, C. Dwork și J. M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography, Proceedings of the 16th International Conference on World Wide Web WWW'07, p. 181-190. (2007)

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/.

[27] Datele genetice sunt un exemplu deosebit de semnificativ de date sensibile care pot fi expuse riscului de reidentificare în cazul în care singurul mecanism de „anonimizare" a acestora este eliminarea identității donatorilor. A se vedea exemplul citat la punctul 2.2.2 de mai sus. A se vedea, de asemenea, John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol. 339, No. 6117 (18 ianuarie 2013), p. 262.

O altă categorie de date care sunt rareori anonimizate doar prin eliminarea identității persoanelor vizate sau prin criptarea parțială a unor atribute este formată din datele de localizare. Modelele de mobilitate ale ființelor umane pot fi suficient de singulare astfel încât partea semantică a datelor de localizare (locurile în care persoana vizată a fost într-un anumit moment), chiar și fără alte atribute, poate să dezvăluie multe caracteristici ale unei persoane vizate[28]. Acest lucru a fost dovedit de mai multe ori în studii academice reprezentative[29].

În acest sens, este necesară avertizarea cu privire la utilizarea pseudonimelor ca modalitate ce permite o protecție adecvată a persoanelor vizate împotriva scurgerilor de identitate sau de atribute. Dacă pseudonimizarea se bazează pe substituția unei identități cu un alt cod unic, prezumția că aceasta constituie o eliminare solidă a posibilității de identificare este naivă și nu ține seama de complexitatea metodelor de identificare și de numeroasele contexte în care acestea pot fi aplicate.

A. 3. „Anonimizarea" prin generalizare

Un exemplu simplu poate contribui la clarificarea abordării bazate pe generalizarea atributelor.

Să analizăm cazul în care un operator de date decide să publice un tabel simplu care conține trei elemente informative sau atribute: un număr de identificare, unic pentru fiecare înregistrare, un element de identificare a locației, care leagă persoana vizată de locul în care locuiește și un element de identificare a proprietății, care indică proprietatea pe care o are persoana vizată; se presupune în continuare că această proprietate este una dintre cele două valori distincte, indicate în mod generic prin {P1, P2}:

Număr de identificare	**Cod de identificare a locației**	**Proprietate**
#1	Roma	P1
#2	Madrid	P1
#3	Londra	P2
#4	Paris	P1
#5	Barcelona	P1
#6	Milano	P2
#7	New York	P2
#8	Berlin	P1

Tabelul A1. Eșantion de persoane vizate reunite în funcție de locație și proprietățile P1 și P2

În cazul în care o persoană, denumită atacator, știe dinainte că o anumită persoană vizată (ținta) care locuiește în Milano este inclusă în tabel, atunci, după analizarea tabelului, acesta poate afla că, întrucât numărul #6 este singura persoană vizată cu respectivul cod de identificare a locației, aceasta deține, de asemenea, proprietatea P2.

[28] Subiectul a fost abordat în unele legislații naționale. De exemplu, în Franța, statisticile publicate privind localizarea sunt anonimizate prin generalizare și permutare. Prin urmare, INSEE publică statistici care sunt generalizate prin agregarea tuturor datelor la o zonă de 40 000 de metri pătrați. Varietatea setului de date este suficientă pentru a păstra utilitatea datelor, iar permutările previn atacurile de inversare a anonimizării în zone răzlețe. În general, agregarea acestei categorii de date și permutarea acestora oferă garanții solide împotriva deducției și a atacurilor de inversare a anonimizării (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. & Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility, Nature. 3, 1376 (2013).

Acest exemplu foarte simplu indică principalele elemente ale oricărei proceduri de identificare aplicate unui set de date care a trecut printr-un așa-zis proces de anonimizare. Și anume, există un atacator care (în mod accidental sau intenționat) a deținut cunoștințe generale privind unele sau toate persoanele vizate dintr-un set de date. Atacatorul vizează crearea unei legături între respectivele cunoștințe generale și datele din setul de date publicat pentru a obține o imagine mai clară a caracteristicilor persoanelor vizate în cauză.

Pentru a face legăturile dintre date și orice tip de cunoștințe generale mai puțin eficiente sau mai puțin imediate, operatorul de date s-ar putea concentra pe codul de identificare a locației, înlocuind orașul în care trăiesc persoanele vizate cu o zonă mai extinsă, cum ar fi țara. Astfel, tabelul ar arăta după cum urmează.

Număr de identificare	**Cod de identificare a locației**	**Proprietate**
#1	Italia	P1
#2	Spania	P1
#3	UK	P2
#4	Franța	P1
#5	Spania	P1
#6	Italia	P2
#7	SUA	P2
#8	Germania	P1

Tabelul A 2. Generalizarea tabelului A1 după naționalitate

Cu această nouă agregare a datelor, cunoștințele generale ale atacatorului cu privire la o persoană vizată identificată (de exemplu, „ținta locuiește în Roma și se află în tabel") nu permit formularea unei concluzii clare cu privire la proprietatea acesteia: aceasta din cauză că cei doi italieni din tabel dețin proprietăți diferite, P1 și, respectiv, P2. Atacatorul rămâne cu o incertitudine de 50 % cu privire la proprietatea entității vizate. Acest exemplu simplu arată efectul generalizării în practica anonimizării. În realitate, în timp ce acest artificiu de generalizare ar putea fi eficient pentru a se reduce la jumătate probabilitatea de a se identifica o țintă italiană, acesta nu este eficient pentru o țintă din alte locații (de exemplu, SUA).

De asemenea, un atacator mai poate afla informații cu privire la ținta spaniolă. În cazul în care cunoștințele generale sunt de tipul „ținta locuiește în Madrid și se află în tabel" sau „ținta locuiește în Barcelona și se află în tabel", atacatorul pot deduce cu o certitudine de 100 % că ținta deține proprietatea P1. Prin urmare, generalizarea nu generează același nivel de confidențialitate sau de rezistență împotriva atacurilor prin deducție pentru întreaga populație din setul de date.

Urmând acest raționament, s-ar putea concluziona că o generalizare mai amplă ar putea fi utilă pentru a se preveni stabilirea oricăror legături – de exemplu, o generalizare în funcție de continent. Astfel, tabelul ar arăta după cum urmează:

Număr de identificare	Cod de identificare a locației	Proprietate
#1	Europa	P1
#2	Europa	P1
#3	Europa	P2
#4	Europa	P1
#5	Europa	P1
#6	Europa	P2
#7	America de Nord	P2
#8	Europa	P1

Tabelul A3. Generalizarea tabelul A1 după continent

Cu acest tip de agregare, toate persoanele vizate din tabel, cu excepția celor care trăiesc în SUA, ar fi protejate împotriva posibilității de a se stabili legături și împotriva atacurilor în vederea identificării, iar orice informații generale de tipul „ținta locuiește în Madrid și se află în tabel" sau „ținta locuiește în Milano și se află în tabel" ar conduce mai degrabă la un anumit nivel de probabilitate în ceea ce privește proprietatea, care se aplică unei anumite persoane vizate (P1 cu o probabilitate de 71,4 % și P2 cu o probabilitate de 28,6 %), decât la stabilirea de legături directe. De asemenea, această generalizare suplimentară este asociată cu o pierdere evidentă și radicală de informații: tabelul nu permite descoperirea potențialelor corelații între proprietăți și locație, și anume, dacă o locație specifică ar putea determina oricare dintre cele două proprietăți cu o probabilitate mai mare, întrucât acesta generează doar așa-numitele distribuții „marginale", și anume, probabilitatea absolută de apariție a proprietății P1 și P2 la nivelul întregii populații (62,5 % și, respectiv, 37,5 %, în exemplul nostru) și pe fiecare continent (astfel cum s-a menționat, 71,4 % și 28,6 % în Europa, 100 % și 0 % în America de Nord).

Exemplul arată, de asemenea, că aplicarea generalizării afectează utilitatea practică a datelor. Unele instrumente de inginerie sunt disponibile în prezent pentru a se stabili în avans (și anume, înainte de publicarea unui set de date) care este nivelul cel mai adecvat de generalizare a atributelor astfel încât să se reducă riscurile de identificare pentru persoanele vizate dintr-un tabel fără a se afecta într-un mod excesiv utilitatea datelor publicate.

k-anonimat

Încercarea de prevenire a atacurilor prin stabilirea de legături, pe baza generalizării atributelor, este cunoscută drept k-anonimat. Această practică este rezultatul unui experiment de reidentificare efectuat la sfârșitul anilor 1990, atunci când o societate americană privată, activă în sectorul sănătății, a publicat o bază de date aparent anonimizată. Anonimizarea a constat în eliminarea numelor persoanelor vizate, dar setul de date a cuprins în continuare date medicale și alte atribute precum codul poștal (codul de identificare a locației unde au locuit persoanele vizate), sexul și data nașterii completă. Aceleași trei elemente {cod poștal, sex, data nașterii completă} au fost incluse, de asemenea, în alte registre publice (de exemplu, lista alegătorilor) și, prin urmare, acestea au putut fi utilizate de către un cercetător din mediul academic pentru a stabili legătura dintre identitatea unor anumite persoane vizate și atributele din setul de date publicat. Cunoștințele generale deținute de atacator (cercetător) ar putea fi după cum urmează: „Știu că persoana vizată din lista de alegători deținând 3 elemente specifice {cod poștal, sex, data nașterii completă} este unică. Există o înregistrare în setul de

date publicat care conține cele trei elemente specifice". S-a observat în mod empiric[30] că marea majoritate (peste 80 %) a persoanelor vizate din registrul public utilizat în acest experiment de cercetare au fost asociate în mod univoc cu o anumită serie de trei elemente, fapt care a făcut posibilă identificarea. Prin urmare, datele nu au fost în mod adecvat anonimizate în acest caz.

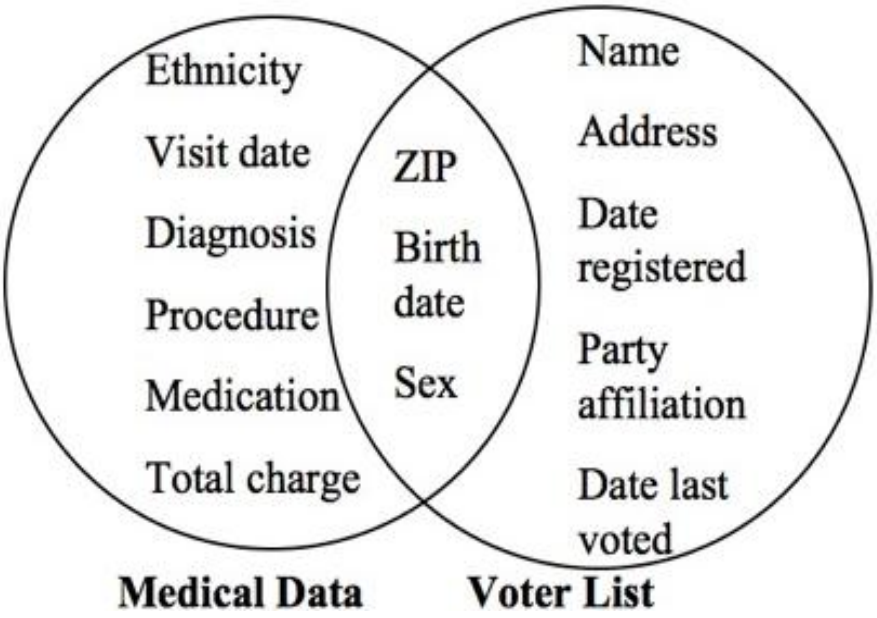

Figura A1. Reidentificare prin stabilirea de legături între date

Legendă:

Etnie		*Adresă*
Data vizitei	*Cod poștal*	*Data înregistrării*
Diagnostic	*Data nașterii*	*Afiliere politică*
Procedură	*Sexul*	*Data ultimei participări la vot*
Medicație		
Tarif total		

Pentru a se reduce eficacitatea unor atacuri similare de stabilire de legături, s-a susținut că operatorii ar trebui să analizeze mai întâi setul de date și să grupeze atributele care ar putea fi utlizate în mod rezonabil de către un atacator pentru a stabili legături între tabelul publicat și o altă sursă auxiliară; fiecare grup ar trebui să includă cel puțin *k* combinații identice de atribute generalizate (și anume, ar trebui să reprezinte o clasă de echivalență a atributelor). Ulterior, seturile de date ar trebui să fie publicate numai după ce au fost divizate în astfel de grupuri omogene. Atributele selectate pentru generalizare sunt cunoscute în literatura de specialitate drept cvasi-identificatori, întrucât cunoașterea acestora ar implica în mod clar identificarea imediată a persoanelor vizate.

Numeroase experimente de identificare au demonstrat neajunsurile unor tabele k-anonimizate ineficient concepute. Acest lucru s-ar putea întâmpla, de exemplu, din cauză că celelalte atribute dintr-o clasă de echivalență sunt identice (astfel cum se întâmplă în clasa de echivalență a persoanelor spaniole vizate din exemplul de la tabelul A2) sau pentru că distribuția acestora este foarte dezechilibrată, cu un grad ridicat de frecvență a unui atribut

[30] L. Sweeney, Weaving Technology and Policy Together to Maintain Confidentiality, Journal of Law, Medicine &Ethics, 25, nos. 2&3 (1997): 98-110.

specific sau în caz contrar, pentru că numărul de înregistrări dintr-o clasă de echivalență este foarte scăzut, permițând în ambele cazuri deducția prin probabilitate, sau pentru că nu există nicio diferență „semantică" semnificativă între atributele clare ale claselor de echivalență (de exemplu, măsurarea cantitativă a unor astfel de atribute ar putea fi în realitate diferită, însă foarte apropiată din punct de vedere numeric sau acestea ar putea aparține unei serii de atribute similare din punct de vedere semantic, de exemplu, aceeași serie a riscului de credit sau aceeași familie de patologii), astfel încât setul de date poate face în continuare vizibile numeroase informații cu privire la persoanele vizate pentru atacuri prin stabilire de legături[31]. Un punct important de notat în acest context este acela că, atunci când datele sunt rare (de exemplu, există puține ocurențe ale unei anumite proprietăți într-o zonă geografică), iar o primă agregare nu este capabilă să grupeze date cu un număr suficient de ocurențe ale diferitelor proprietăți (de exemplu, mai rămâne un număr mic de ocurențe ale câtorva proprietăți care pot fi localizate într-o anumită zonă geografică), este necesară o agregare suplimentară a atributelor pentru a se realiza anonimizarea vizată.

l-diversitatea

Pe baza acestor observații au fost propuse de-a lungul anilor variante ale k-anonimatului și au fost elaborate o serie de criterii tehnice în scopul de a se consolida practica anonimizării prin generalizare, astfel încât să se reducă riscurile de atacuri prin creare de legături. Acestea sunt bazate pe proprietățile probabilistice ale seturilor de date. Mai precis, s-a adăugat o constrângere suplimentară, și anume, ca fiecare atribut dintr-o clasă de echivalență să apară de cel puțin *l* ori, astfel încât un atacator să rămână întotdeauna cu un grad semnificativ de incertitudine cu privire la atribute, inclusiv în prezența unor cunoștințe generale cu privire la o anumită persoană vizată. Acest lucru este echivalent cu a afirma că un set de date (sau diviziune a acestuia) ar trebui să dețină un număr minim de ocurențe ale unei proprietăți selectate: acest artificiu ar putea reduce riscul de reidentificare. Acesta este obiectivul aplicării anonimizării cu l-diversitate. Un exemplu de astfel de aplicare este prezentat în tabelele A4 (datele originale) și A5 (rezultatul prelucrării). Evident, prin prelucrarea adecvată a codului de identificare a locației și a vârstei persoanelor din tabelul A4, generalizarea atributelor determină o creștere substanțială a gradului de incertitudine cu privire la atributele reale ale fiecărei persoane vizate din sondaj. De exemplu, chiar dacă atacatorul știe că o persoană vizată face parte din prima clasă de echivalență, el nu poate stabili mai departe dacă o persoană deține proprietatea X, Y sau Z, întrucât în clasa respectivă (și în orice altă clasă de echivalență) există cel puțin o înregistrare care prezintă astfel de proprietăți.

[31] Trebuie subliniat faptul că se pot stabili corelații inclusiv după ce înregistrările datelor au fost grupate în funcție de atribute. Atunci când operatorul de date cunoaște tipurile de corelații pe care dorește să le verifice, acesta poate selecta atributele cele mai relevante. De exemplu, rezultatele sondajului PEW nu fac obiectul unor atacuri detaliate prin deducție, de aceea sunt în continuare foarte utile pentru identificarea unor corelații între datele demografice și interese (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx).

Număr de identificare	Cod de identificare a locației	Vârsta	Proprietate
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tabelul A4. Un tabel cu persoane fizice grupate în funcție de locație, vârstă și trei proprietăți X, Y și Z

Număr de identificare	Cod de identificare a locației	Vârsta	Proprietate
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Tabelul A5. Un exemplu de versiune 1-diversificată a tabelului A4

t-apropiere:

Cazul specific al atributelor din cadrul unei diviziuni, care sunt distribuite în mod inegal sau care aparțin unei game mici de valori sau semnificații semantice este tratat prin abordarea denumită *t-apropiere*. Aceasta reprezintă o îmbunătățire suplimentară a anonimizării prin generalizare și constă în practica aranjării datelor în așa fel încât să se obțină clase de echivalență de natură să reflecte cât mai mult posibil distribuția inițială a atributelor în setul de date original. În acest scop, se utilizează o procedură în două etape, în esență după cum urmează. Tabelul A6 reprezintă baza de date originală care include înregistrări clare ale persoanelor vizate, grupate în funcție de locație, vârstă, salariu și două categorii de proprietăți similare, din punct de vedere semantic, și anume, (X1, X2, X3) și (Y1, Y2, Y3) (de exemplu, clase similare de risc de credit, boli similare). În primul rând, tabelul este *l-diversificat* cu *l*=1 (tabelul A7), prin gruparea înregistrărilor în clase de echivalență similare din punct de vedere semantic și anonimizare vizată deficitară; ulterior acesta este prelucrat în vederea obținerii t-apropierii (tabelul A8) și a unui grad mai ridicat de variabilitate în cadrul fiecărei diviziuni. În fapt, în cea de a doua etapă, fiecare clasă de echivalență include înregistrări din ambele categorii de proprietăți. Trebuie menționat că vârsta și codul de identificare a locației au particularități diferite în diferitele etape ale procesului: acest lucru înseamnă că fiecare atribut

ar putea necesita criterii diferite de generalizare pentru a se obține anonimizarea vizată, iar acest lucru presupune, la rândul său, o inginerie specifică și o sarcină de calcul adecvată din partea operatorilor de date.

Număr de identificare	Cod de identificare a locației	Vârstă	Salariu	Proprietate
1	1127	29	30K	X1
2	1112	22	32K	X2
3	1128	27	35K	X3
4	1215	43	50K	X2
5	1219	52	120K	Y1
6	1216	47	60K	Y2
7	1115	30	55K	Y2
8	1123	36	100K	Y3
9	1117	32	110K	X3

Tabelul A6. Un tabel cu persoane fizice grupate în funcție de locație, vârstă, salariu și două categorii de proprietăți

Număr de identificare	Cod de identificare a locației	Vârstă	Salariu	Proprietate
1	11**	2*	30K	X1
2	11**	2*	32K	X2
3	11**	2*	35K	X3
4	121*	>40	50K	X2
5	121*	>40	120K	Y1
6	121*	>40	60K	Y2
7	11**	3*	55K	Y2
8	11**	3*	100K	Y3
9	11**	3*	110K	X3

Tabelul A 7. O versiune *l*-diversificată a tabelului A6

Număr de identificare	Cod de identificare a locației	Vârstă	Salariu	Proprietate
1	112*	<40	30K	X1
3	112*	<40	35K	X3
8	112*	<40	100K	Y3
4	121*	>40	50K	X2
5	121*	>40	120K	Y1
6	121*	>40	60K	Y2
2	111*	<40	32K	X2
7	111*	<40	55K	Y2
9	111*	<40	110K	X3

Tabelul A8. O versiune *t*-apropiată a tabelului A6

Trebuie să se precizeze în mod clar că obiectivul generalizării atributelor persoanelor vizate în astfel de modalități studiate poate fi uneori realizat numai pentru un număr mic de înregistrări,

și nu pentru integralitatea acestora. Bunele practici ar trebui să asigure faptul că fiecare clasă de echivalență conține mai multe persoane și că nu mai pot fi posibile atacuri prin deducție. În orice caz, această abordare necesită o evaluare aprofundată a datelor disponibile de către operatorii de date împreună cu o evaluare combinatorie a diferitelor alternative (de exemplu, cu serii diferite de amplitudine, particularități diferite ale locației sau vârstei etc.). Cu alte cuvinte, anonimizarea prin generalizare nu poate fi rezultatul unei prime încercări brute a operatorilor de date de a substitui valorile analitice ale atributelor dintr-o înregistrare cu intervale, întrucât sunt necesare mai multe abordări cantitative specifice, cum ar fi evaluarea nivelului de entropie a atributelor din cadrul fiecărei diviziuni sau măsurarea distanței dintre distribuțiile atributelor originale și distribuția din fiecare clasă de echivalență.