ČLANAK 29. RADNA SKUPINA ZA ZAŠTITU PODATAKA

0829/14/HR
WP216

Mišljenje 05/2014 o tehnikama anonimizacije

doneseno 10. travnja 2014.

RADNA SKUPINA ZA ZAŠTITU POJEDINACA U VEZI S OBRADOM OSOBNIH PODATAKA

uspostavljena Direktivom 95/46/EZ Europskog parlamenta i Vijeća od 24. listopada 1995.,

uzimajući u obzir njezine članke 29. i 30.,

uzimajući u obzir njezin poslovnik,

DONIJELA JE OVO MIŠLJENJE:

SAŽETAK

U ovom mišljenju radna skupina analizira učinkovitost i ograničenja postojećih tehnika anonimizacije u odnosu na pravnu pozadinu EU-a u području zaštite podataka i daje preporuke za postupanje s ovim tehnikama uzimajući u obzir rezidualni rizik utvrđivanja identiteta svojstven svakoj od njih.

Radna skupina potvrđuje potencijalnu vrijednost anonimizacije posebno kao strategiju iskorištavanja prednosti „otvorenih podataka" za pojedince i društvo u cjelini uz ublažavanje rizika za dotične pojedince. Međutim, na temelju rezultata studije slučaja i objavljenih istraživanja može se zaključiti da je teško kreirati zaista anoniman skup podataka i pritom zadržati što više osnovnih podataka potrebnih za obavljanje zadaće.

S obzirom na Direktivu 95/46/EZ i ostale mjerodavne pravne instrumente EU-a, anonimizacija je rezultat obrade osobnih podataka kako bi se nepovratno spriječilo utvrđivanje identiteta. Pritom bi nadzornici podataka trebali uzeti u obzir nekoliko elemenata, uzimajući u obzir sva sredstva koja će se „vjerojatno razumno" koristiti za utvrđivanje identiteta (od strane nadzornika ili bilo koje treće stranke).

Anonimizacija predstavlja daljnju obradu osobnih podataka i, kao takva, mora zadovoljiti uvjet sukladnosti uzimajući u obzir pravne temelje i okolnosti daljnje obrade. Osim toga, anonimizirani podaci ne nalaze se u području primjene zakonodavstva o zaštiti podataka, ali osobe čiji se podaci obrađuju i dalje mogu imati pravo na zaštitu prema ostalim odredbama (poput onih koje štite povjerljivost komunikacija).

U ovom su mišljenju opisane glavne tehnike anonimizacije, odnosno, randomizacija i generalizacija. U mišljenju se posebno raspravlja o dodavanju šuma, permutaciji, diferencijalnoj privatnosti, sažimanju, k-anonimnosti, l-raznolikosti i t-bliskosti. Objašnjuju se njihova načela, prednosti i slabosti te uobičajene pogreške i propusti povezani s korištenjem svake tehnike.

U mišljenju se razrađuje pouzdanost svake tehnike na temelju tri kriterija:

i. je li još uvijek moguće izdvojiti pojedinca,

ii. je li još uvijek moguće povezati zapise koji se odnose na pojedinca i

iii. mogu li se izvesti zaključci o pojedincu?

Poznavanje glavnih prednosti i slabosti svake tehnike pomaže u odabiru načina oblikovanja prikladnog procesa anonimizacije u danom kontekstu.

Obrađuje se i pseudonimizacija radi razjašnjavanja određenih zamki i pogrešnih predodžaba: pseudonimizacija nije metoda anonimizacije. Njome se samo smanjuje povezivost skupa podataka s originalnim identitetom osobe čiji se podaci obrađuju te je s time u skladu korisna sigurnosna mjera.

Mišljenje završava zaključkom da se tehnikama anonimizacije mogu pružiti jamstva privatnosti i one se mogu koristiti za generiranje učinkovitih procesa anonimizacije, ali samo ako je njihova primjena primjereno organizirana – što znači da preduvjeti (kontekst) i cilj/ciljevi procesa anonimizacije moraju biti jasno utvrđeni kako bi se postigla ciljana anonimizacija uz prikupljanje određenih korisnih podataka. O optimalnom rješenju treba

odlučiti na temelju pojedinačnog slučaja, po mogućnosti koristeći kombinaciju različitih tehnika, uzimajući u obzir praktične preporuke razvijene u ovom mišljenju.

Konačno, nadzornici podataka trebaju uzeti u obzir da anonimizirani skup podataka i dalje može predstavljati rezidualne rizike za osobe čiji se podaci obrađuju. S jedne strane, anonimizacija i ponovno utvrđivanje identiteta su zaista aktivna područja istraživanja i redovito se objavljuju nova otkrića, ali s druge se strane, čak i anonimizirani podaci, poput statistika, mogu koristiti za obogaćivanje profila pojedinaca i tako stvoriti nova pitanja u zaštiti podataka. Zbog tog razloga anonimizaciju ne bi trebalo smatrati jednokratnom vježbom i nadzornici podataka trebaju redovito ponovno procjenjivati očekivane rizike.

1 Uvod

Iako se uređajima, senzorima i mrežama stvaraju velike količine i nove vrste podataka, a trošak pohranjivanja podataka postaje zanemariv, postoji sve veći javni interes i potražnja za ponovnim korištenjem tih podataka. „Otvoreni podaci" mogu pružiti očite koristi za društvo, pojedince i organizacije, ali samo ako se poštuju svačija prava na zaštitu njihovih osobnih podataka i privatnog života.

Anonimizacija može biti dobra strategija za očuvanje prednosti i ublažavanje rizika. Kad je skup podataka zaista anonimiziran i više nije moguće utvrditi identitet pojedinaca, više se ne primjenjuje europsko pravo o zaštiti podataka. Međutim, na temelju studija slučaja i istraživačkih publikacija može se zaključiti da kreiranje zaista anonimnog skupa podataka iz bogatog skupa osobnih podataka, pri čemu se zadržava što više osnovnih podataka potrebnih za zadatak, nije jednostavan pothvat. Na primjer, skup podataka koji se smatra anonimnim može se kombinirati s nekim drugim skupom podataka na način da je moguće utvrditi identitet jednog ili više pojedinaca.

Radna skupina u ovom mišljenju analizira učinkovitost i ograničenja postojećih tehnika anonimizacije u odnosu na pravnu pozadinu EU-a u području zaštite podataka i daje preporuke za pažljivo i odgovorno korištenje ovih tehnika s ciljem stvaranja procesa anonimizacije.

2 Definicije i pravna analiza

2.1. Definicije u pravnom kontekstu EU-a

U uvodnoj izjavi (26) Direktive 95/46/EZ poziva se na anonimizaciju radi isključivanja anonimiziranih podataka iz područja primjene zakonodavstva o zaštiti podataka:

„budući da se načela zaštite moraju primjenjivati na sve podatke u vezi s utvrđenim osobama ili osobama koje se mogu utvrditi; budući da je, kako bi se utvrdilo može li se osobu utvrditi, potrebno uzeti u obzir sva sredstva koja nadzornik ili bilo koja druga osoba može opravdano koristiti da utvrdi navedenu osobu; budući da se načela zaštite ne primjenjuju na podatke koji su anonimni na takav način da se osoba čiji se podaci obrađuju više ne može utvrditi; budući da pravila ponašanja u smislu članka 27. ove Direktive mogu biti korisni instrument za davanje uputa u vezi s načinom na koji se podaci mogu napraviti anonimnima i zadržati u obliku u kojem utvrđivanje identiteta osobe čiji se podaci obrađuju više nije moguća;".[1]

Iz tumačenja uvodne izjave (26) može se dobiti konceptualna definicija anonimizacije. Uvodna izjava (26) znači da za anonimizaciju svih podataka iz podataka treba ukloniti suvišne elemente tako da više nije moguće utvrditi identitet osobe čiji se podaci obrađuju. Drugim riječima, ti se podaci moraju obraditi na način da ih više nije moguće koristiti za utvrđivanje identiteta fizičke osobe korištenjem „svih sredstava koja opravdano mogu koristiti" nadzornik ili treća stranka. Važan je čimbenik da obrada mora biti nepovratna. Direktivom se ne objašnjava kako bi se trebao ili mogao provoditi takav proces ponovnog utvrđivanja

[1] Osim toga treba napomenuti da je ovaj pristup korišten i u nacrtu uredbe EU-a o zaštiti podataka, u uvodnoj izjavi (23) „za utvrđivanje je li moguće utvrditi identitet osobe treba uzeti u obzir sva sredstva koja opravdano mogu koristiti nadzornik ili bilo koja druga osoba za utvrđivanje identiteta pojedinca".

identiteta[2]. Naglasak je na rezultatu: podaci trebaju biti takvi da ne dopuštaju utvrđivanje identiteta osobe čiji se podaci obrađuju putem „svih" „mogućih" i „opravdanih" sredstava. Upućuje se na pravila ponašanja kao alat za utvrđivanje mogućih mehanizama anonimizacije te zadržavanje u obliku u kojem utvrđivanje identiteta osobe čiji se podaci obrađuju „više nije moguće". U Direktivi se jasno utvrđuje visok standard.

U Direktivi o e-privatnosti (Direktiva 2002/58/EZ) također se na vrlo sličan način govori o „anonimizaciji" i „anonimnim podacima". U uvodnoj izjavi (26) navodi se sljedeće:

> *„Podaci o prometu koji su se koristili za marketinške komunikacijske usluge ili za pružanje usluga s posebnom tarifom također se trebaju brisati ili učiniti anonimnima nakon pružanja usluge.".*

U skladu s time, u članku 6. stavku 1. navodi se sljedeće:

> *„Podaci o prometu koji se odnose na pretplatnike i korisnike i koje je davatelj javne komunikacijske mreže ili javno dostupne elektroničke komunikacijske usluge obradio i pohranio moraju se obrisati ili učiniti anonimnima kada više nisu potrebni u svrhu prijenosa komunikacije, ne dovodeći u pitanje stavke 2., 3. i 5. ovog članka te članak 15. stavak 1."*
>
> U članku 9. stavku 1. navodi se, između ostalog, sljedeće:
> *„Ako se mogu obraditi podaci o lokaciji koji nisu podaci o prometu, koji se odnose na korisnike ili pretplatnike javnih komunikacijskih mreža ili javno dostupnih elektroničkih komunikacijskih usluga, takvi podaci mogu se obraditi samo nakon što su učinjeni anonimnima, odnosno uz pristanak korisnika ili pretplatnika, u mjeri i u trajanju potrebnom za pružanje usluge s dodatnom vrijednosti."*

Osnovno obrazloženje jest da bi rezultat anonimizacije kao tehnike primijenjene na osobne podatke trebao, u sadašnjem stanju tehnologije, biti jednako trajan kao brisanje, odnosno, onemogućivati obradu osobnih podataka.[3]

2.2. Pravna analiza

Na temelju analize teksta povezanog s anonimizacijom u vodećim instrumentima EU-a o zaštiti podataka moguće je istaknuti četiri ključne značajke:

- Anonimizacija može biti rezultat obrade osobnih podataka s ciljem nepovratnog sprečavanja utvrđivanja identiteta osobe čiji se podaci obrađuju.

- Može se predvidjeti nekoliko tehnika anonimizacije, nema propisane norme u zakonodavstvu EU-a.

---

[2] Ovaj je koncept detaljnije razrađen na str. 8 ovog mišljenja.

[3] Treba podsjetiti da se anonimizacija u međunarodnim normama također definira kao ISO 29100 – „Proces kojim se podaci kojima se utvrđuje identitet osobe (PII) nepovratno izmjenjuju na način da nadzornik PII-a sam ili u srudanji s bilo kojom drugom strankom više ne mogu izravno ili neizravno utvrditi PII" (ISO 29100:2011). Nepovratnost izmjene osobnih podataka za omogućivanje izravnog ili neizravnog utvrđivanja identiteta je ključ i za ISO. S ovog stajališta postoji značajna usklađenost s načelima i konceptima iz Direktive 95/46. To se također primjenjuje na definicije koje se nalaze u nekim nacionalnim zakonima (na primjer u Italiji, Njemačkoj i Sloveniji) gdje je naglasak na neutvrdivosti identiteta i upućuje se na „neproporcionalni trud" ponovnog utvrđivanja identiteta (Njemačka, Slovenija). Međutim, francuski zakon o zaštiti podataka predviđa da podaci ostaju osobni podaci čak i ako je izuzetno teško i nevjerojatno ponovno utvrditi identitet osobe čiji se podaci obrađuju – dakle nema odredbe koja se odnosi na test „opravdanosti".

- Važnost treba pridati kontekstnim elementima: treba uzeti u obzir „sva vjerojatno opravdana" sredstva koja će nadzornik i treće stranke koristiti za utvrđivanje identiteta, pridajući posebnu pozornost onome što je u zadnje vrijeme postalo, u trenutačnom stanju tehnologije, „vjerojatno razumno" (s obzirom na povećanje računalne snage i dostupnih alata).

- Anonimizaciji je svojstven faktor rizika: taj faktor rizika treba uzeti u obzir prilikom procjene valjanosti svake tehnike anonimizacije – uključujući moguće upotrebe bilo kojih podataka koji su „učinjeni anonimnima" pomoću takve tehnike – a treba procijeniti i ozbiljnost i vjerojatnost ovog rizika.

U ovom se mišljenju radije koristi pojam „tehnika anonimizacije", nego „anonimnost" ili „anonimni podaci" za isticanje svojstvenog rezidualnog rizika od ponovnog utvrđivanja identiteta povezanog s bilo kojom tehničkom i organizacijskom mjerom s ciljem da se podaci učine „anonimnima".

2.2.1. Zakonitost procesa anonimizacije

Prvo, anonimizacija je tehnika koja se primjenjuje na osobne podatke za postizanje nepovratnog ponovnog utvrđivanja identiteta. Stoga je početna pretpostavka takva da je osobne podatke trebalo prikupiti i obraditi u skladu s mjerodavnim zakonodavstvom o zadržavanju podataka u obliku koji je moguće utvrditi.

U ovom je kontekstu proces anonimizacije, dakle obrada takvih osobnih podataka kako bi se postigla njihova anonimnost, stupanj „daljnje obrade". Kao takva, ova obrada mora biti skladu s testom kompatibilnosti iz smjernica radne skupine navedenih u mišljenju 03/2013 o ograničavanju svrhe[4].

To znači da se, u načelu, pravna osnova za anonimizaciju može pronaći u bilo kojem od razloga spomenutih u članku 7. (uključujući zakoniti interes nadzornika) pod uvjetom da se također poštuju zahtjevi o kvaliteti podataka iz članka 6. Direktive i uz primjereno uzimanje u obzir posebnih okolnosti i svih čimbenika navedenih u mišljenju radne skupine o ograničavanju svrhe[5].

S druge strane treba napomenuti odredbe sadržane u članku 6. stavku 1. točki (e) Direktive 95/46/EZ (ali i u članku 6. stavku 1. i članku 9. stavku 1. Direktive o e-privatnosti) jer prikazuju potrebu za čuvanjem osobnih podataka „u obliku koji omogućuje utvrđivanje identiteta" ne dulje nego što je potrebno za potrebe prikupljanja ili daljnje obrade podataka.

Ovom se odredbom upućuje na to da osobni podaci trebaju biti najmanje anonimizirani „automatski" (podložno različitim pravnim zahtjevima, kao što su oni spomenuti u Direktivi o e-privatnosti u pogledu podataka o prometu). Ako nadzornik podataka želi zadržati takve

[4] Mišljenje 03/2013 radne skupine iz članka 29. za zaštitu pojedinaca u vezi s obradom osobnih podataka, dostupno na: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5] To posebno znači da treba provesti temeljitu procjenu s obzirom na sve mjerodavne okolnosti, osobito one koje se odnose na sljedeće ključne čimbenike:
a) odnos između svrha prikupljanja osobnih podataka i svrha daljnje obrade;
b) kontekst u kojem su prikupljani osobni podaci i razumna očekivanja od osoba čiji se podaci obrađuju u pogledu daljnjeg korištenja podataka;
c) prirodu osobnih podataka i učinak daljnje obrade na osobe čiji se podaci obrađuju;
d) zaštitne mjere koje je donio nadzornik za osiguranje pravedne obrade i sprečavanje nepotrebnog učinka na osobe čiji se podaci obrađuju.

osobne podatke nakon postizanja svrhe originalne ili daljnje obrade, trebalo bi koristiti tehnike anonimizacije za nepovratno sprečavanje utvrđivanja identiteta.

S time u skladu radna skupina smatra da se anonimizacija kao način daljnje obrade osobnih podataka može smatrati sukladnom s originalnim svrhama obrade, ali samo pod uvjetom da je proces anonimizacije takav da se njime pouzdano mogu proizvesti anonimizirane podatke u smislu opisanom u ovom dokumentu.

Također treba naglasiti da anonimizacija treba biti u skladu s pravnim ograničenjima na koja je podsjetio Sud Europske unije u svojoj odluci u slučaju C-553/07 (*College van burgemeester en wethouders van Rotterdam protiv M.E.E. Rijkeboer*), o potrebi zadržavanja podataka u obliku koji se može utvrditi kako bi se, na primjer, osobama čiji se podaci obrađuju moglo omogućiti ostvarivanje prava na pristup podacima. Sud EU-a je presudio da „*Države članice moraju u skladu s člankom 12. točkom (a) Direktive [95/46] osigurati pravo pristupa podacima u pogledu primatelja ili kategorija primatelja osobnih podataka i sadržaja objavljenih podataka ne samo što se tiče sadašnjosti, već i prošlosti. Države članice moraju odrediti rok za pohranu tih podataka i pružiti pristup tim podacima što predstavlja pravednu ravnotežu između interesa osobe čiji se podaci obrađuju da zaštiti svoju privatnost s jedne strane, posebno njegovim pravom na prigovor i korištenje pravnih sredstava, i s druge strane teret koji za nadzornika predstavlja obveza pohranjivanja tih podataka.*"

Ovo je posebno mjerodavno u slučaju kada se nadzornik podataka oslanja na članak 7. točku (f) Direktive 95/46 u pogledu anonimizacije: opravdani interes nadzornika podataka uvijek mora biti u skladu s pravima i temeljnim slobodama osoba čiji se podaci obrađuju.

Na primjer, na temelju rezultata istrage nizozemskog DPA u 2012.-2013. o korištenju DPI tehnologija od strane četiri mobilna operatora pravni temelj utvrđuje se prema članku 7. točki (f) Direktive 95/46 za anonimizaciju sadržaja podataka o prometu što prije nakon prikupljanja tih podataka. U članku 6. Direktive o e-privatnosti zaista se propisuje da se podaci o prometu pretplatnika i korisnika koje je obradio i pohranio davatelj javne komunikacijske mreže ili javno dostupne elektroničke komunikacijske usluge moraju obrisati ili učiniti anonimnima što je prije moguće. U ovom slučaju postoji odgovarajući pravni temelj u članku 7. Direktive o zaštiti podataka zato što je dopušten prema članku 6. Direktive o e-privatnosti. To bi se također moglo predstaviti na drugi način: ako vrsta obrade podataka nije dopuštena prema članku 6. Direktive o e-privatnosti, ne može biti pravnog temelja u članku 7. Direktive o zaštiti podataka.

2.2.2. Mogućnost potencijalnog utvrđivanja anonimiziranih podataka

U Mišljenju 4/2007 o osobnim podacima radna skupina je detaljno obradila koncept osobnih podataka, fokusirajući se na elemente definicije iz članka 2. točke (a) Direktive 95/46/EZ uključujući ovaj dio te definicije: „koja se može utvrditi ili čiji identitet se može utvrditi". U ovom kontekstu radna je skupina također zaključila da bi „anonimizirani podaci stoga bili anonimni podaci koji su se ranije odnosili na osobu čiji se identitet može utvrditi, ali kada to utvrđivanje više nije moguće".

Radna skupina je stoga već objasnila da se u Direktivi predlaže provjera „sredstava … koja treba opravdano koristiti" kao kriterij koji se primjenjuje kako bi se procijenilo je li proces anonimizacije dovoljno stabilan, tj. je li utvrđivanje identiteta postalo „opravdano" nemoguće. Poseban kontekst i okolnosti pojedinog slučaja imaju izravan učinak na mogućnost utvrđivanja identiteta. U tehničkom prilogu ovom mišljenju dana je analiza učinka odabira najprikladnije tehnike.

Kao što je već naglašeno, istraživanja, alati i računalna snaga se razvijaju. Stoga nije moguće ni korisno predvidjeti iscrpno navođenje okolnosti kada utvrđivanje identiteta više nije moguće. Međutim, treba uzeti u obzir i prikazati neke ključne čimbenike.

Prvo, može se tvrditi da bi se nadzornici podataka trebali fokusirati na konkretna sredstva koja bi bila potrebna za poništavanje tehnike anonimizacije, posebno u pogledu troškova i znanja i iskustva potrebnih za provedbu tih sredstava i procjene njihove vjerojatnosti i ozbiljnosti. Na primjer, trebali bi uravnotežiti svoje napore i troškove anonimizacije (i u pogledu vremena i potrebnih sredstava) u odnosu na sve veću dostupnost povoljnih tehničkih sredstava za utvrđivanje identiteta pojedinaca u skupovima podataka, sve veću javnu dostupnost ostalih skupova podataka (poput onih koji su omogućeni u vezi s politikama ‚otvorenih podataka'), i mnoge primjere nepotpune anonimizacije koja dovodi do naknadnih nepovoljnih, ponekad nenadoknadivih učinaka na osobe čiji se podaci obrađuju.[6] Treba napomenuti da se rizik utvrđivanja identiteta s vremenom povećava i također ovisi o razvoju informacijske i komunikacijske tehnologije. Pravni propisi, ako postoje, stoga trebaju biti formulirani tehnološki neutralno i u idealnom slučaju uzimati u obzir promjene u razvojnim potencijalima informacijske tehnologije.[7]

Drugo, „sredstva koja će se vjerojatno opravdano koristiti za utvrđivanje je li moguće utvrditi identitet osobe" ona su koja koristi „nadzornik ili bilo koja druga osoba". Stoga je važno shvatiti da kada nadzornik podataka ne obriše originalne podatke (koji se mogu utvrditi) na razini slučaja i nadzornik podataka preda dio ovog skupa podataka (na primjer nakon uklanjanja ili maskiranja podataka koji se mogu utvrditi), rezultirajući skup podataka i dalje čini osobne podake. Samo ako nadzornik podataka skupi podatke na razinu kada se pojedinačni slučajevi više ne mogu utvrditi, rezultirajući skup podataka može se kvalificirati kao anoniman. Na primjer: ako neka organizacija prikuplja podatke o pojedinačnim putovanjima, pojedinačni obrasci putovanja na razini slučaja i dalje bi se kvalificirali kao osobni podaci za bilo koju stranku, dok god nadzornik podataka (ili bilo koja druga stranka) i dalje ima pristup originalnim podacima, čak i ako su iz skupa pruženog trećim strankama uklonjeni izravni identifikatori. Ali ako bi nadzornik podataka obrisao neobrađene podatke i samo omogućio skupne statističke podatke trećim strankama na visokoj razini, poput ‚ponedjeljkom na ruti X ima 160 % više putnika nego utorkom', oni bi se kvalificirali kao anonimni podaci.

Učinkovitim rješenjem anonimizacije onemogućuje se strankama da izdvajaju pojedinca u skupu podataka, od povezivanja dva zapisa unutar skupa podataka (ili između dva odvojena skupa podataka) i od izvođenja bilo kakvih podataka u takvom skupu podataka. Općenito govoreći, uklanjanje elemanata koji izravno utvrđuju identitet stoga samo po sebi nije dovoljno kako bi se osiguralo da više nije moguće utvrđivanje identiteta osobe čiji se podaci obrađuju. Često će biti potrebno poduzeti dodatne mjere za sprečavanje utvrđivanja identiteta, još jednom ovisno o kontekstu i svrhama obrade za koje su namijenjeni anonimizirani podaci.

[6] Zanimljivo, u izmjenama Nacrta uredbe o zaštiti općih podataka koje je nedavno podnio Europski parlament (21. listopada 2013.) u uvodnoj izjavi (23) posebno se navodi da „Za utvrđivanje hoće li se sredstva vjerojatno opravdano koristiti za utvrđivanje identiteta pojedinca, treba uzeti u obzir sve objektivne čimbenike, poput troškova i količine vremena potrebne za utvrđivanje identiteta, uzimajući u obzir i dostupnu tehnologiju u vrijeme obrade i tehnološki razvoj".

[7] Vidjeti mišljenje 4/2007 radne skupine iz članka 29. za zaštitu pojedinaca u vezi s obradom osobnih podataka, str. 15.

PRIMJER:
Profili genetskih podataka primjer su osobnih podataka koji mogu biti rizični za utvrđivanje identiteta ako je jedina tehnika koja se koristi uklanjanje identiteta donatora zbog jedinstvene prirode određenih profila. Već je pokazano u literaturi[8] da kombinacija javno dostupnih genetičkih resursa (npr. genealoške evidencije, nekrološke evidencije, rezultati upita iz pretraživača) i metapodataka o darivateljima DNK (vrijeme donacije, dob, mjesto boravišta) mogu otkriti identitet određenih pojedinaca čak i ako je DNK darovan „anonimno".

Obje obitelji tehnika anonimizacije – randomizacija i generalizacija podataka –[9] imaju nedostatke; međutim, svaka od njih može u danim okolnostima i kontekstu biti primjerena za ostvarivanje željene svrhe bez ugrožavanja privatnosti osobe čiji se podaci obrađuju. Treba biti jasno da ‚utvrđivanje identiteta' ne znači samo mogućnost pronalaska imena i/ili adrese osobe, već uključuje i potencijalnu mogućnost utvrđivanja identiteta izdvajanjem, povezivošću i izvođenjem zaključaka. Nadalje, kako bi pravo o zaštiti podataka bilo mjerodavno, nije važno koje su namjere nadzornika podataka ili primatelja. Dok god je podatke moguće utvrditi, primjenjuju se pravila o zaštiti podataka.

Ako treća stranka obrađuje skup podataka obrađen tehnikom anonimizacije (originalni nadzornik podataka ih je učinio anonimnima i objavio), ona to može učiniti zakonito bez potrebe za uzimanjem u obzir zahtjeve o zaštiti podataka pod uvjetom da ne može (izravno ili neizravno) utvrditi identitet osoba čiji se podaci obrađuju u originalnom skupu podataka. Međutim, treće stranke moraju uzimati u obzir sve iznad navedene čimbenike u odnosu na kontekst i okolnosti (uključujući posebna obilježja tehnika anonimizacije kako ih je primijenio originalni nadzornik podataka) u odlučivanju kako koristiti i, posebno, kombinirati takve anonimizirane podatke za njihove svrhe – jer rezultati i posljedice mogu dovesti do različitih vrsta odgovornosti s njihove strane. Ako su ti čimbenici i obilježja takvi da dovode do neprihvatljivog rizika od utvrđivanja identiteta osoba čiji se podaci obrađuju, obrada će još jednom dospjeti u područje primjene zakona o zaštiti podataka.

Gornji popis ni u kojem slučaju nije iscrpan, već mu je svrha pružiti opće smjernice o pristupu procjeni potencijala utvrđivanja identiteta određenog skupa podataka koji je podvrgnut anonimizaciji u skladu s različitim dostupnim tehnikama. Svi gore navedeni čimbenici mogu se smatrati mnogim faktorima rizika koje nadzornici podataka trebaju ocijeniti u anonimiziranim skupovima podataka i treće stranke prilikom korištenja tih „anonimiziranih" skupova podataka u vlastite svrhe.

2.2.3. Rizici korištenja anonimiziranih podataka

Prilikom razmatranja o korištenju tehnika anonimizacije, nadzornici podataka trebaju uzeti u obzir sljedeće rizike:

- Posebna je zamka smatrati da su pseudonimizirani podaci istovjetni anonimiziranima. U odjeljku o tehničkoj analizi bit će objašnjeno da se pseudonimizirani podaci ne mogu izjednačiti s anonimiziranim podacima jer i dalje omogućuju izdvajanje pojedine osobe čiji se podaci obrađuju i povezivi su preko različitih skupova podataka. Pseudonimnost će vjerojatno omogućiti utvrđivanje identiteta i stoga ostaje u području primjene pravnog režima zaštite

[8] Vidjeti John Bohannon, Genealoške baze podataka omogućuju imenovanje anonimnih donatora DNK, Science, Svezak 339, br. 6117 (18. siječnja 2013.), str. 262.
[9] Osnovna svojstva i razlike između ove dvije tehnike anonimizacije opisani su u odjeljku 3 ispod („Tehnička analiza").

podataka. To je posebno mjerodavno u kontekstu znanstvenog, statističkog ili povijesnog istraživanja.[10]

PRIMJER:
Tipičan primjer pogrešne predodžbe o pseudonimizaciji dobro je poznata nezgoda korporacije AOL (America On Line). Godine 2006. javno je objavljena baza podataka koja sadržava dvadeset milijuna ključnih riječi za pretraživanje više od 650 000 korisnika u tromjesečnom razdoblju, s jedinom mjerom za očuvanje privatnosti koja se sastojala od zamjene korisničkog imena AOL-a numeričkim atributom. To je dovelo do javnog utvrđivanja i lociranja nekih korisnika. Pseudonimizirani nizovi upita pretraživača, posebno ako su upareni s ostalim atributima, poput IP adresa ili ostalih parametara konfiguracije klijenta, imaju vrlo visoku moć utvrđivanja.

- Druga je pogreška smatrati da se ispravno anonimiziranim podacima (koji su zadovoljili sve gore navedene uvjete i kriterije i koji su po definiciji izvan područja primjene Direktive o zaštiti podataka) pojedinci lišavaju zaštitne mjere – u prvom redu jer se na korištenje tih podataka mogu primjenjivati drugi dijelovi zakonodavstva. Na primjer, člankom 5. stavkom 3. Direktive o e-privatnosti onemogućuje se pohranjivanje i pristup „podacima" bilo koje vrste (uključujući neosobne podatke) na terminalnoj opremi bez pristanka pretplatnika/korisnika jer je to dio šireg načela povjerljivosti komunikacija.

- Treći primjer nepažnje također je posljedica nerazmatranja utjecaja ispravno anonimimiziranih podataka na pojedince, u određenim okolnostima, posebno u slučaju izrade profila. Područje privatnog života pojedinca zaštićeno je člankom 8. Europske konvencije o ljudskim pravima i člankom 7. Povelje EU-a o temeljnim pravima; kao takvi, čak iako se zakoni o zaštiti podataka više ne primjenjuju na ovu vrstu podataka, korištenje anonimiziranih skupova podataka i objavljenih za korištenje trećim strankama može izazvati gubitak privatnosti. Posebna je pozornost potrebna prilikom rukovanja anonimiziranim podacima posebno kadgod se ti podaci koriste (često u kombinaciji s ostalim podacima) za donošenje odluka koje imaju utjecaj (pa makar i neizravan) na pojedince. Kao što je radna skupina već spomenula u ovom mišljenju i posebno objasnila u mišljenju o konceptu „ograničavanja svrhe" (Mišljenje 03/2013)[11], legitimna očekivanja osoba čiji se podaci obrađuju o daljnjoj obradi njihovih podataka treba procijeniti s obzirom na mjerodavne čimbenike povezane s kontekstom – kao što je priroda odnosa između osoba čiji se podaci obrađuju i nadzornika podataka, primjenjivih pravnih obveza, transparentnosti postupaka obrade.

3 Tehnička analiza, stabilnost tehnologija i tipične pogreške

Postoje različite prakse i tehnike anonimizacije s promjenjivim stupnjevima pouzdanosti. U ovom će se odjeljku obraditi osnovne točke koje trebaju razmotriti nadzornici podataka u njihovoj primjeni posebno uzimajući u obzir jamstvo koje može postići određena tehnika s obzirom na trenutačno stanje tehnologije i uzimajući u obzir tri rizika koji su ključni za anonimizaciju:

- *izdvajanje* što odgovara mogućnosti izoliranja nekoliko ili svih zapisa koji utvrđuju identitet pojedinca u skupu podataka;
- *povezivost* koja predstavlja mogućnost povezivanja najmanje dva zapisa o istoj osobi čiji se podaci obrađuju ili skupini osoba čiji se podaci obrađuju (ili u istoj

[10] Vidjeti također Mišljenje 4/2007 radne skupine iz članka 29. za zaštitu pojedinaca u vezi s obradom osobnih podataka, str. 18-20.
[11] Dostupno na http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

bazi podataka ili u dvije različite baze podataka). Ako napadač može utvrditi (npr. pomoću korelacijske analize) da su dva zapisa dodijeljena istoj skupini pojedinaca ali ne može izdvojiti pojedince u ovoj grupi, tehnika pruža otpor od „izdvajanja" ali ne od povezivosti;
- *izvođenje zaključaka* koje predstavlja mogućnost zaključivanja sa značajnom vjerojatnošću, vrijednosti atributa iz vrijednosti skupa ostalih atributa.

Stoga bi rješenje za ova tri rizika pouzdano štitilo od ponovnog utvrđivanja identiteta provedenog najvjerojatnijim i najopravdanijim sredstvima koje nadzornik podataka i bilo koja treća stranka mogu koristiti. S ovime u vezi radna skupina naglašava da su tehnike deidentifikacije i anonimizacije predmet tekućeg istraživanja i takvo je istraživanje dosljedno pokazalo da nijedna tehnika nije bez nedostataka. Općenito govoreći postoje dva različita pristupa anonimizaciji: prvi se temelji na ***randomizaciji*** dok se drugi temelji na ***generalizaciji***. U mišljenju se također obrađuju drugi koncepti kao što su *pseudonimizacija, diferencijalna privatnost, l-raznolikost, t-bliskost.*

U ovom se mišljenju koristi sljedeći rječnik u ovom odjeljku: skup podataka sastoji se od različitih zapisa o pojedincima (osobama čiji se podaci obrađuju). Svaki se zapis odnosi na jednu osobu čiji se podaci obrađuju i sastoji se od skupa vrijednosti (ili „unosa", npr.: 2013.) za svaki atribut (npr. godina). Skup podataka je zbirka zapisa koji se na drugi način mogu oblikovati kao tablica (ili skup tablica) ili kao označeni/težinski grafikon, što je sve češći slučaj danas. Primjeri u mišljenju odnosit će se na tablice, ali se primjenjuju i na ostale grafičke prikaze zapisa. Kombinacije atributa povezanih s osobom čiji se podaci obrađuju ili skupinom osoba čiji se podaci obrađuju mogu se nazivati kvazi-identifikatorima. U nekim slučajevima skup podataka može imati višestruke zapise o istom pojedincu. „Napadač" je treća stranka (tj. ni nadzornik podataka ni obrađivač podataka) koji pristupa originalnim zapisima bilo slučajno ili namjerno.

3.1. Randomizacija

Randomizacija je obitelj tehnika kojom se mijenja istinitost podataka kako bi se uklonila snažna poveznica između podataka i pojedinca. Ako su podaci dovoljno nesigurni, onda više ne mogu upućivati na određenu osobu. Randomizacija sama po sebi neće smanjiti osebujnost svakog zapisa jer će se svaki zapis i dalje izvoditi iz pojedine osobe čiji se podaci obrađuju, ali može štititi od napada/rizika od izvođenja zaključaka i može se kombinirati s tehnikama generalizacije kako bi se pružila jača jamstva privatnosti. Mogu se koristiti dodatne tehnike kako bi se osiguralo da pojedinac ne može utvrditi zapis.

3.1.1. Dodavanje šuma

Tehnika dodavanja šuma posebno je korisna kada atributi mogu imati važan nepovoljan učinak na pojedince i sastoji se od modifikacije atributa u skupu podataka tako da budu manje precizni uz zadržavanje opće distribucije. Kod obrade skupa podataka promatrač će pretpostaviti da su vrijednosti precizne, ali to će biti istinito samo u određenoj mjeri. Na primjer, ako je visina pojedinca originalno izmjerena do najbližeg centimetra, anonimizirani skup podataka može sadržavati visinu koja je precizna do samo +-10cm. Ako se ova tehnika primjenjuje učinkovito, treća stranka neće moći utvrditi identitet pojedinca niti bi mogla popraviti podatke ili na drugi način otkriti kako su podaci izmijenjeni.

Dodavanje šuma obično će trebati kombinirati s ostalim tehnikama anonimizacije kao što je uklanjanje očitih atributa i kvazi-identifikatora. Razina buke treba ovisiti o potrebi razine potrebnih podataka i utjecaja na privatnost pojedinca kao rezultat objave zaštićenih atributa.

3.1.1.1. Jamstva

- *Izdvajanje:* Još je moguće izdvojiti zapise pojedinca (možda na način da se ne može utvrditi identitet) čak i ako su zapisi manje pouzdani.
- *Povezivost:* Još je moguće povezati zapise istog pojedinca, ali zapisi su manje pouzdani i stoga se stvarni zapis može povezati s umjetno dodanim (npr. s „šumom"). U nekim slučajevima nepravilna atribucija može izložiti osobu čiji se podaci obrađuju značajnom i čak visokom riziku u odnosu na pravilnu.
- *Izvođenje zaključaka:* Napadi izvođenja mogu biti mogući ali stopa uspjeha bit će niža i vjerojatni su neki lažni pozitivni (i lažni negativni odgovori).

3.1.1.2. Uobičajene pogreške

- Dodavanje promjenjivog šuma: Ako šum nije semantički održiv (tj. „disproporcionalan" i ne poštuje logiku između atributa u skupu), tada će napadač koji ima pristup bazi podataka moći filtrirati šum i u nekim slučajevima ponovno generirani nedostajuće unose. Nadalje, ako je skup podataka previše raštrkan[12], postoji mogućnost povezivanja unosa podataka koji sadrže šum s vanjskim izvorom.
- Pretpostavljajući da je dodavanje šuma dovoljno: dodavanje šuma dopunska je mjera kojom se napadaču otežava pristup osobnim podacima. Osim ako je šum veći od podataka sadržanih u skupu podataka, ne bi se trebalo pretpostaviti da dodavanje šuma predstavlja samostalno rješenje za anonimizaciju.

3.1.1.3. Propusti dodavanja šuma

Vrlo poznati eksperiment ponovnog utvrđivanja identiteta jest onaj proveden na bazi podataka korisnika pružatelja video sadržaja Netflix. Znanstvenici su analizirali geometrijska svojstva te baze podataka koja sadržava više od 100 milijuna ocjena u rasponu od $1-5$ za preko 18 000 filmova koje je ocijenilo skoro 500 000 korisnika. Društvo ju je javno objavilo nakon „anonimizacije" u skladu s internom politikom zaštite privatnosti pri čemu su uklonjeni svi podaci kojima se utvrđuje identitet korisnika osim ocjena i datuma. Dodan je šum jer su ocjene neznatno povećane ili smanjene.

Unatoč tome, utvrđeno je da se 99 % korisničkih zapisa može nedvojbeno utvrditi u skupu podataka uz korištenje 8 ocjena i datuma s 14-dnevnim pogreškama kao kriterijima odabira, dok je smanjenje kriterija odabira (2 ocjene i 3-dnevna pogreška) i dalje omogućivalo utvrđivanje identiteta 68 % korisnika.[13]

3.1.2. Permutacija

Ova tehnika, koja se sastoji od miješanja vrijednosti atributa u tablici tako da su neki od njih umjetno povezani s različitim osobama čiji se podaci obrađuju, korisna je kada je važno zadržati točnu distribuciju svakog atributa unutar skupa podataka.

Permutacija se može smatrati posebnim oblikom dodavanja šuma. U klasičnoj tehnici šuma atributi se modificiraju s nasumičnim vrijednostima. Stvaranje konstantnog šuma može biti težak zadatak i lagana modifikacija vrijednosti ne mora pružiti prikladnu privatnost. Kao

[12] Ovaj koncept je dalje razrađen u prilogu na str. 30.
[13] Narayanan A. i Šmatikov V. (svibanj 2008.). Pouzdana deanonimizacija velikih raštrkanih skupova podataka. U *Sigurnost i privatnost, 2008. SP 2008. Simpozij IEEE-a* (str. 111. – 125.). IEEE.

alternativa, tehnike permutacije mijenjaju vrijednosti unutar skupa podataka tako da ih samo prebacuju iz jednog zapisa u drugi. Takvim će se zamjenama osigurati da raspon i distribucija vrijednosti ostanu jednaki, a korelacije između vrijednosti i pojedinaca ne. Ako dva ili više atributa imaju logički odnos ili statističku korelaciju i neovisno se permutiraju, takav će se odnos uništiti. Stoga može biti važno permutirati skup povezanih atributa tako da se ne uništi logički odnos, inače bi napadač mogao utvrditi permutirane atribute i poništiti permutaciju.

Na primjer, ako podskup atributa u medicinskoj bazi podataka smatramo „razlozima za hospitalizaciju/simptomima/nadležnim odjelom", snažan logički odnos će povezati vrijednosti u većini slučajevima, a permutacija samo jedne vrijednosti bi stoga bila otkrivena i čak bi se mogla poništiti.

Slično dodavanju šuma, permutacija možda sama po sebi ne omogućuje anonimizaciju i uvijek je treba kombinirati s uklanjanjem očitih atributa/kvazi-identifikatora.

3.1.2.1. Jamstva

- *Izdvajanje:* Kao s dodavanjem šuma, još je moguće izdvojiti zapise pojedinca ali zapisi su manje pouzdani.
- *Povezivost:* Ako permutacija utječe na atribute i kvazi-identifikatore, njome se može onemogućiti „ispravno" povezivanje atributa interno i eksterno sa skupom podataka, ali omogućuje se „neispravna" povezivost, jer se realni unos može povezati s različitom osobom čiji se podaci obrađuju.
- *Izvođenje zaključaka:* Još uvijek je moguće izvući zaključke iz skupa podataka, posebno ako su atributi povezani ili imaju snažne logičke odnose; međutim, kada nije poznato koji su atributi permutirani, napadač treba uzeti u obzir da se njegovo izvođenje zaključaka temelji na pogrešnoj hipotezi i stoga ostaje moguće samo probabilističko izvođenje zaključaka.

3.1.2.2. Uobičajene pogreške

- Izbor pogrešnog atributa: permutacija neosjetljivih ili nerizičnih atributa ne bi rezultirala značajnim dobitkom u pogledu zaštite osobnih podataka. Ako bi se naime osljetljivi/rizični atributi i dalje povezivali s originalnim atributom, tada bi napadač i dalje mogao izvesti osjetljive podatke o pojedincima.
- Nasumična permutacija atributa: Ako su dva atributa snažno povezana, tada nasumična permutacija atributa neće pružiti snažna jamstva. Ova uobičajena pogreška prikazana je u Tablici 1.
- Pretpostavka da je permutacija dovoljna: Kao i u slučaju dodavanja šuma, samom permutacijom ne omogućuje se anonimnost i treba je kombinirati s ostalim tehnikama poput uklanjanja očitih atributa.

3.1.2.3. Propusti permutacije

Ovim se primjerom prikazuje kako nasumična permutacija atributa donosi slabo jamstvo privatnosti kada postoje logičke poveznice između različitih atributa. Slijedeći pokušaj anonimizacije, jednostavno je izvesti zaključak o dohotku svakog pojedinca ovisno o radnom mjestu (i godini rođenja). Na primjer, izravnim pregledom podataka može se zaključiti da je generalni direktor u tablici vrlo vjerojatno rođen 1957. i ima najveću plaću, dok je nezaposlena osoba rođena 1964. i ima najmanji dohodak.

Godina	Spol	Radno mjesto	Dohodak (permutiran)
1957.	M	Inženjer	70 000
1957.	M	Generalni direktor	5 000
1957.	M	Nezaposlena osoba	43 000
1964.	M	Inženjer	100 000
1964.	M	Direktor	45 000

Tablica 1. Neučinkoviti primjer anonimizacije permutacijom povezanih atributa

3.1.3. Diferencijalna privatnost

Diferencijalna privatnost[14] dio je obitelji tehnika randomizacije s drugačijim pristupom: dok umetanje šuma zapravo dolazi prethodno kada se očekuje objava skupa podataka, diferencijalna privatnost može se koristiti kada nadzornik podataka generira anonimizirane poglede skupa podataka uz zadržavanje kopije originalnih podataka. Takvi anonimizirani pogledi obično bi se mogli generirati kroz podskup upita za pojedinu treću stranku. Podskup uključuje određenu količinu naknadno namjerno dodanog nasumičnog šuma. Diferencijalna privatnost govori nadzorniku podataka koliko šuma treba dodati, i u kojem obliku, da dobije potrebna jamstva privatnosti.[15] U tom će kontekstu biti posebno važno neprekidno pratiti (barem za svaki novi upit) bilo koju mogućnost utvrđivanja identiteta pojedinca u skupu rezultata upita. Međutim treba objasniti da tehnike diferencijalne privatnosti neće promijeniti originalne podatke i stoga, dok god ostaju originalni podaci, nadzornik podataka može utvrditi identitet pojedinaca u rezultatima upita diferencijalne privatnosti uzimajući u obzir sva sredstva koja će se vjerojatno opravdano koristiti. Takve rezultate dakle treba smatrati osobnim podacima.

Jedna korist pristupa koji se temelji na diferencijalnoj privatnosti jest u činjenici da se skupovi podataka pružaju ovlaštenim trećim strankama kao odgovor na poseban upit radije nego kroz objavu pojedinačnog skupa podataka. Da pomogne u reviziji nadzornik podataka može zadržati popis svih upita i zahtjeva osiguravajući da treće stranke ne pristupaju podacima za koje nisu ovlašteni. Upit se također može podvrgnuti tehnikama anonimizacije uključujući dodavanje šuma ili zamjenu radi daljnje zaštite privatnosti. Još je uvijek otvoreno pitanje istraživanja kako pronaći dobar mehanizam odgovora na upite koji u isto vrijeme može prilično točno odgovoriti na bilo koja pitanja (znači s manje šumova) uz istovremeno očuvanje privatnosti.

Za ograničavanje napada izvođenja zaključaka i povezivosti potrebno je pratiti upite koje je izdao subjekt i promatrati podatke dobivene o osobama čiji se podaci obrađuju; s time u skladu, baze podataka „diferencijalne privatnosti" ne bi trebalo koristiti na besplatnim pretraživačima koji ne nude sljedivost subjekata koji postavljaju upite.

[14] Dwork, C. (2006.). Difrencijalna privatnost. U *Automata, languages and programming* (str. 1. – 12.). Springer Berlin Heidelberg.
[15] Cf. Ed Felten (2012.) Zaštita privatnosti dodavanjem šuma. URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

3.1.3.1 Jamstva

- *Izdvajanje:* Ako se prikazuju samo statistički podaci i dobro su odabrana pravila koja se primjenjuju na skup, ne bi trebalo biti moguće koristiti odgovore za izdvajanje pojedinca.
- *Povezivost:* Korištenjem višestrukih zahtjeva postoji mogućnost povezivanja unosa povezanih s određenim pojedincem između dva odgovora.
- *Izvođenje zaključaka:* Moguće je izvesti zaključke o podacima o pojedincima ili skupinama koristeći višestruke zahtjeve.

3.1.3.2. Uobičajene pogreške

- Neubacivanje dovoljno šuma: Za sprečavanje povezivanja s prethodnim znanjem, izazov je pružiti minimalni dokaz o tome je li određena osoba čiji se podaci obrađuju ili skupina osoba čiji se podaci obrađuju doprinijela skupu podataka. Najveća teškoća iz perspektive zaštite podataka jest moći generirati odgovarajuću količinu šuma koji treba dodati istinitim odgovorima kako bi se zaštitila privatnost pojedinca i istovremeno očuvala korisnost objavljenih odgovora.

3.1.3.3 Propusti diferencijalne privatnosti

Neovisno postupanje sa svakim upitom: Kombinacijom rezultata upita može se omogućiti objavu podataka koji bi trebali biti tajni. Ako povijest upita nije zadržana, tada napadač može konstruirati višestruka pitanja za bazu podataka „diferencijalne privatnosti" koji progresivno smanjuju amplitudu izlaznog uzorka dok se ne pojavi posebno svojstvo pojedinačne osobe čiji se podaci obrađuju ili skupine osoba čiji se podaci obrađuju, deterministički ili s vrlo velikom vjerojatnošću. Nadalje, dodatno ograničenje jest izbjegavati pogrešku u mišljenju da su podaci anonimni za treću stranku, dok nadzornik podataka još može utvrditi identitet osobe čiji se podaci obrađuju u originalnoj bazi podataka uzimajući u obzir sva sredstva koja će se vjerojatno opravdano koristiti.

3.2. Generalizacija

Generalizacija je druga obitelj tehnika anonimizacije. Ovaj pristup sastoji se od generalizacije, ili razrjeđivanja, atributa osoba čiji se podaci obrađuju putem izmjene odgovarajućeg raspona ili reda veličina (tj. regije prije nego grada, mjeseca prije nego tjedna). Iako generalizacija može biti učinkovita za sprečavanje izdvajanja, njome se ne omogućuje učinkovita anonimizacija u svim slučajevima; posebno zahtijeva specifične i sofisticirane kvantitativne pristupe za sprečavanje povezivosti i izvođenja zaključaka.

3.2.1. Sažimanje i K-anonimnost

Tehnike sažimanja i K-anonimnosti imaju za cilj spriječiti izdvajanje osobe čiji se podaci obrađuju njihovim grupiranjem s najmanje *k* ostalih pojedinaca. Kako bi se to postiglo, vrijednosti atributa se generaliziraju do takve mjere da svaki pojedinac dijeli istu vrijednost. Na primjer, smanjivanjem rascjepkanosti lokacije od grada do zemlje uključen je veći broj osoba čiji se podaci obrađuju. Pojedinačni datumi rođenja mogu se generalizirati u raspon datuma ili grupirati po mjesecu ili godini. Ostali numerički atributi (npr. plaće, težina, visina ili doza lijeka) mogu se generalizirati po vrijednostima intervala (npr. plaća 20 000 – 30 000 EUR). Te se metode mogu koristiti kada korelacija preciznih vrijednosti atributa može kreirati kvazi-identifikatore.

3.2.1.1. Jamstva

- *Izdvajanje:* Budući da iste atribute sada dijeli *k* korisnika, više ne bi trebalo biti moguće izdvojiti pojedinca unutar grupe *k* korisnika.
- *Povezivost:* Iako je povezivost ograničena, postoji mogućnost povezivanja zapisa po grupama *k* korisnika. Zatim je, unutar ove skupine, vjerojatnost da dva zapisa odgovaraju istim pseudo-identifikatorima *1/k* (što može biti značajno više od vjerojatnosti da su takvi unosi nepovezivi).
- *Izvođenje zaključaka:* Glavni nedostatak modela k-anonimnosti jest da se njime ne onemogućuje nikakvu vrstu napada izvođenja zaključaka. Ako su doista svi *k* pojedinci unutar iste skupine, tada je jednostavno dohvatiti vrijednost ovog svojstva ako je poznato kojoj pojedinac pripada skupini.

3.2.1.2. Uobičajene pogreške

- Neki kvazi-identifikatori nedostaju: Kritičan parametar kad se smatra da je k-anonimnost prag od *k*. Što je veća vrijednost *k*, jača su jamstva privatnosti. Uobičajena je pogreška umjetno povećati vrijednost *k* smanjenjem razmatranog skupa kvazi-identifikatora. Smanjenjem kvazi-identifikatora olakšava se stvaranje klastera k-korisnika zbog inherentne moći utvrđivanja povezane s ostalim atributima (posebno ako su neki od njih osjetljivi ili imaju vrlo veliku entropiju, kao što je slučaj s vrlo rijetkom atributima). Kritična je pogreška ne razmotriti sve kvazi-identifikatore prilikom odabira atributa za generalizaciju; ako se neki atributi mogu koristiti za izdvajanje pojedinca u klasteru *k*, tada generalizacijom nije moguće zaštititi neke pojedince (vidjeti primjer u tablici 2.).
- Mala vrijednost *k*: Ciljanje male vrijednosti *k* predstavlja sličan problem. Ako je *k* premali, ponder bilo kojeg pojedinca u klasteru previše je značajan i napadi izvođenja zaključaka imaju veću stopu uspjeha. Na primjer ako je k = 2, tada je vjerojatnost da dva pojedinca dijele isto svojstvo veća nego kada je k > 10.
- Negrupiranje pojedinaca s istim ponderom: Grupiranje skupa pojedinaca s nejednakom distribucijom atributa također može biti problematično. Učinak pojedinčevog zapisa u skupu podataka će varirati: neki će predstavljati značajan udio za unose dok doprinosi ostalih ostaju prilično beznačajni. Stoga je važno osigurati da *k* bude dovoljno velik da pojedinci ne predstavljaju prevažan udio unosa u klasteru.

3.1.3.3. Propusti k-anonimnosti

Glavni problem u pogledu k-anonimnosti jest da se njome ne sprečavaju napadi izvođenja zaključaka. U sljedećem primjeru ako napadač zna da je određeni pojedinac u skupu podataka i da je rođen 1964., također zna da je pojedinac imao srčani udar. Nadalje, ako znamo da je ovaj skup podataka dobiven od francuske organizacije, onda svaki pojedinac živi u Parizu jer su prve tri brojke pariškog poštanskog broja 750*).

Godina	Spol	Poštanski broj	Dijagnoza
1957.	M	750*	Srčani udar
1957.	M	750*	Kolesterol
1957.	M	750*	Kolesterol
1964.	M	750*	Srčani udar
1964.	M	750*	Srčani udar

Tablica 2. Primjer loše izvedene k-anonimizacije

3.2.2. L-raznolikost/T-bliskost

L-raznolikost proširuje k-anonimnost kako bi osigurala da više nisu mogući deterministički napadi izvođenja zaključaka tako da u svakoj klasi ekvivalencije svaki atribut ima najmanje *l* različitih vrijednosti.

Osnovni cilj koji treba postići jest ograničiti pojavljivanje klasa ekvivalencije sa slabom varijabilnosti atributa, tako da napadač s prethodnim znanjem o određenoj osobi čiji se podaci obrađuju uvijek bude ostavljen sa značajnom nesigurnosti.

L-raznolikost je korisna za zaštitu podataka od napada izvođenja zaključaka kada su vrijednosti atributa dobro distribuirane. Međutim treba naglasiti da se ovom tehnikom ne može spriječiti curenje podataka ako su atributi unutar podpodjele nejednako distribuirani ili pripadaju malom rasponu vrijednosti ili semantičkih značenja. Na kraju, l-raznolikost je podložna probabilističkim napadima izvođenja zaključaka.

T-bliskost je pročišćavanje l-raznolikosti tako da ima za cilj kreirati klase ekvivalencije koje sliče početnoj distribuciji atributa u tablici. Ova je tehnika korisna kada je važno držati podatke što je bliže moguće originalnima; u tu se svrhu na klasu ekvivalencije stavlja daljnje ograničenje, naime takvo da u svakoj klasi ekvivalencije ne bi trebalo postojati samo najmanje *l* različitih vrijednosti, već i da je svaka vrijednost zastupljena onoliko puta koliko je potrebno da zrcali početnu distribuciju svakog atributa.

3.2.2.1. Jamstva

- *Izdvajanje:* Kao k-anonimnost, l-raznolikost i t-bliskost mogu osigurati da se zapisi povezani s pojedincem ne mogu izdvojiti u bazi podataka.
- *Povezivost:* l-raznolikost i t-bliskost nisu poboljšanje u odnosu na k- anonimnost s obzirom na nepovezivost. Pitanje je isto kao i u bilo kojem klasteru: vjerojatnost da isti unosi pripadaju istoj osobi čiji se podaci obrađuju veći od 1/N (pri čemu je N broj osoba čiji se podaci obrađuju u bazi podataka).
- *Izvođenje zaključaka:* Glavno poboljšanje l-raznolikosti i t-bliskosti nad k-anonimnosti jest da više nije moguće sa 100 %-tnom pouzdanosti organizirati napade izvođenja zaključaka na ‚l-raznoliku' ili ‚t-blisku' bazu podataka.

3.2.2.2. Uobičajene pogreške

- Zaštita vrijednosti osjetljivih atributa miješanjem s ostalim osjetljivim atributima: Nije dovoljno imati dvije vrijednosti atributa u klasteru kako bi se pružila jamstva privatnosti. Zapravo, distribucija osjetljivih vrijednosti u svakom klasteru trebala bi sličiti distribuciji tih vrijednosti u ukupnoj populaciji, ili bi barem trebala biti jednaka u klasteru.

3.2.2.3. Propusti l-raznolikosti

U donjoj tablici l-raznolikost odobrena je s obzirom na atribut „Dijagnoza”; međutim, znajući da je pojedinac rođen 1964. u ovoj tablici, još je s vrlo velikom vjerojatnošću moguće pretpostaviti da je imao srčani udar.

Godina	Spol	Poštanski broj	Dijagnoza
1957.	M	750*	**Srčani udar**
1957.	M	750*	**Kolesterol**
1957.	M	750*	**Kolesterol**
1957.	M	750*	**Kolesterol**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Kolesterol**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**
1964.	M	750*	**Srčani udar**

Tablica 3. L-raznolika tablica u kojoj vrijednosti „Dijagnoze” nisu jednako distribuirane

Prezime	Datum rođenja	Spol
Smith	1964.	M
Rossi	1964.	M
Dupont	1964.	M
Jansen	1964.	M
Garcia	1964.	M

Tablica 4. Znajući da su ovi pojedinci u tablici 3., napadač bi mogao zaključiti da su imali srčani udar

4. Pseudonimizacija

Pseudonimizacija se sastoji od zamjene jednog atributa (obično jedinog atributa) u zapisu drugim. Stoga je i dalje vjerojatno moguće neizravno utvrditi identitet fizičke osobe; s time u skladu, kada se pseudonimizacija koristi sama, neće rezultirati anonimnim skupom podataka. Ipak, o njoj se u ovom mišljenju raspravlja zbog mnogih pogrešnih predodžaba i pogrešaka u vezi s njezinim korištenjem.

Pseudonimizacijom se smanjuje povezivost skupa podataka s originalnim identitetom osobe čiji se podaci obrađuju; kao takva je korisna mjera sigurnosti ali ne i metoda anonimizacije.

Rezultat pseudonimizacije može biti neovisan od početne vrijednosti (kao što je slučaj s nasumičnim brojem koji je generirao nadzornik ili prezimenom koje je odabrala osoba čiji se podaci obrađuju) ili se može izvesti iz originalnih vrijednosti atributa ili skupa atributa npr. funkcija raspršivanja ili enkripcijska shema.

Najčešće se koriste sljedeće tehnike pseudonimizacije:

- Enkripcija s tajnim ključem: u ovom slučaju nositelj ključa može jednostavno ponovno utvrditi identitet svake osobe čiji se podaci obrađuju dekripcijom skupa podataka, iako u kodiranom obliku. Pod pretpostavkom da je primijenjena najmodernija enkripcijska shema, dekripcija je moguća samo uz poznavanje ključa.
- Funkcija raspršivanja: ovo odgovara funkciji kojom se vraća izlaz fiksne veličine iz ulaza bilo koje veličine (unos može biti pojedinačni atribut ili skup atributa) i ne može se poništiti; to znači da rizik od poništavanja, uočen u slučaju enkripcije, više ne postoji. Međutim, ako je poznat raspon vrijednosti unosa i funkcija raspršivanja, mogu se ponoviti kroz funkciju raspršivanja kako bi se izvela točna vrijednost za pojedini zapis. Na primjer, ako je skup podataka pseudonimiziran raspršivanjem nacionalnog identifikacijskog broja, tada se ovo može izvesti jednostavnim raspršivanjem svih mogućih vrijednosti unosa i usporedbom rezultata s tim vrijednostima u skupu podataka. Funkcije raspršivanja su obično oblikovane da se mogu relativno brzo proračunati i podložne su brutalnim napadima.[16] Prethodno proračunate tablice također se mogu kreirati da omoguće masovno poništavanje velikog skupa raspršenih vrijednosti.

  Korištenjem slane funkcije raspršivanja (kada se nasumična vrijednost, poznata kao „sol", dodaje atributu koji se raspršuje) može se smanjiti vjerojatnost izvođenja vrijednosti unosa ali ipak, proračun originalne vrijednosti atributa skrivene iza rezultata slane funkcije raspršivanja još može biti izvediv opravdanim sredstvima.[17]
- Funkcija raspršivanja ključem s pohranjenim ključem: to odgovara posebnoj funkciji raspršivanja koja koristi tajni ključ kao dodatni unos (ovo se razlikuje od slane funkcije raspršivanja jer sol obično nije tajna). Nadzornik podataka može ponoviti funkciju atributa koristeći tajni ključ, ali napadaču je puno teže ponoviti funkciju bez poznavanja ključa jer je broj mogućnosti za testiranje dovoljno velik da bude nepraktičan.

[16] Takvi se napadi sastoje u isprobavanju svih vjerojatnih unosa kako bi se kreirale tablice podudarnosti.
[17] Posebno ako je poznata vrsta atributa (ime, broj socijalnog osiguranja, datum rođenja itd.) Za dodavanje računalnog zahtjeva, može se pouzdati u funkciju raspršivanja derivacijom ključa, kada je izračunata vrijednost nekoliko puta raspršena s malo soli.

- Deterministička enkripcija ili funkcija raspršivanja ključem s brisanjem ključa: ova se tehnika može izjednačiti s odabirom nasumičnog broja kao pseudonima za svaki atribut u bazi podataka i zatim obrisati tablicu podudarnosti. Ovim se rješenjem omogućuje[18] smanjenje rizika povezivosti između osobnih podataka u skupu podataka i onih koji se odnose na istog pojedinca u drugom skupu podataka ako se koristi drugačiji pseudonim. Uzimajući u obzir najnoviji algoritam, napadaču će biti računski teško dekriptirati ili ponoviti funkciju, jer bi to impliciralo testiranje svakog mogućeg ključa, pod uvjetom da ključ nije dostupan.
- Tokenizacija: ova se tehnika obično primjenjuje u (čak i ako nije ograničena na njega) financijskom sektoru ili za zamjenu identifikacijskijh brojeva kartice vrijednostima koji su smanjili korisnost za napadača. Izvodi se iz prethodnih koji se obično temelje na primjeni jednosmjernih enkripcijskih mehanizama ili dodjeli, pomoću funkcije indeksa, rednog broja ili nasumce generiranog broja koji nije matematički izveden iz originalnih podataka.

4.1. Jamstva

- *Izdvajanje:* Još je moguće izdvojiti zapise pojedinaca jer se pojedinac i dalje utvrđuje jedinim atributom koji je rezultat funkcije pseudonimizacije (= pseudonimizirani atribut).
- *Povezivost:* Povezivost će i dalje biti uobičajena između zapisa koji koriste isti pseudonimizirani atribut za upućivanje na istog pojedinca. Čak i ako se za istu osobu čiji se podaci obrađuju koriste različiti pseudonimizirani atributi, povezivost je i dalje moguća pomoću drugih atributa. Samo ako se u skupu podataka ne može koristiti drugi atribut za utvrđivanje identiteta osobe čiji se podaci obrađuju i ako je uklonjena svaka poveznica između originalnog atributa i pseudonimiziranog atributa (uključujući brisanjem originalnih podataka), neće biti očitog unakrsnog upućivanja između dva skupa podataka koristeći različite pseudonimizirane atribute.
- *Izvođenje zaključaka:* Napadi izvođenja zaključaka na stvarni identitet osobe čiji se podaci obrađuju mogući su unutar skupa podataka ili kroz različite baze podataka koje koriste isti pseudonimizirani atribut za pojedinca, ili ako su razumljivi sami po sebi i ne maskiraju pravilno originalni identitet osobe čiji se podaci obrađuju.

4.2. Uobičajene pogreške

- Vjerovanje da je pseudonimizirani skup podataka anonimiziran: Nadzornici podataka često pretpostavljaju da je uklanjanje ili zamjena jednog ili više atributa dovoljno da se skup podataka učini anonimnim. Iz mnogih se primjera može zaključiti da to nije tako; jednostavnom izmjenom identifikatora ne sprečava se nekoga da utvrdi identitet osobe čiji se podaci obrađuju ako u skupu podataka ostanu kvazi-identifikatori, ili ako vrijednosti ostalih atributa i dalje mogu utvrditi identitet pojedinca. U mnogim slučajevima može biti jednostavno utvrditi identitet pojedinca u pseudonimiziranom skupu podataka kao i s originalnim podacima. Trebalo bi poduzeti dodatne korake da se skup podataka smatra anonimiziranim, uključujući uklanjanje i generalizaciju atributa ili brisanje originalnih podataka ili barem njihova dovođenja na visoko sažetu razinu.
- Uobičajene pogreške pri korištenju pseudonimizacije kao tehnike za smanjenje povezivosti:

[18] Ovisno o ostalim atributima u skupu podataka i o brisanju originalnih podataka.

- Korištenje istog ključa u različtim bazama podataka: eliminacija povezivosti različitih skupova podataka jako ovisi o korištenju algoritma s ključem i o činjenici da će jedan pojedinac odgovarati različitim pseudonimiziranim atributima u različitim kontekstima. Stoga je važno izbjegavati korištenje istog ključa u različitim bazama podataka kako bi se mogla smanjiti povezivost.
- Korištenje različitih ključeva („rotirajućih ključeva") za različite korisnike: može biti privlačno korištenje različitih ključeva za različite skupove korisnika i mijenjanje ključa na temelju korištenja (na primjer, korištenje istog ključa za bilježenje 10 unosa koji se odnose na istog korisnika). Međutim, ako se dobro ne organizira, ovom se operacijom može izazvati pojavljivanje uzoraka i djelomično smanjiti planirane prednosti. Na primjer, rotiranje ključa pomoću posebnih pravila za određene pojedince olakšalo bi povezivost unosa koji odgovaraju navedenim pojedincima. Također, nestanak periodičnog pseudonimiziranog podatka u bazi podataka u vrijeme kada se pojavljuje novi podatak može signalizirati da se oba zapisa odnose na istu fizičku osobu.
- Zadržavanje ključa: ako se tajni ključ pohranjuje uz pseudonimizirane podatke, a podaci su ugroženi, tada napadač ima mogućnost jednostavno povezati pseudonimizirane podatke s njihovim originalnim atributom. Isto se primjenjuje ako se ključ pohranjuje odvojeno od podataka, ali na nesiguran način.

4.3. Nedostaci pseudonimizacije

- *Zdravstvena zaštita*

1. **Ime, adresa, datum rođenja**	**2.** **Razdoblje davanja za posebnu pomoć**	**3.** **Indeks tjelesne mase**	**6.** **Referentni broj skupine istraživanja**
[redacted]	< 2 godine	15	QA5FRD4
	> 5 godina	14	2B48HFG
	< 2 godine	16	RC3URPQ
	> 5 godina	18	SD289K9
	< 2 godine	20	5E1FL7Q

Tablica 5. Primjer pseudonimizacije raspršivanjem (ime, adresa, datum rođenja) koje se lako može poništiti

Kreiran je skup podataka za ispitivanje odnosa između težine osobe i primitka davanja za posebnu pomoć. Originalni skup podataka uključivao je ime, adresu i datum rođenja osobe čiji se podaci obrađuju, ali to je obrisano. Referentni broj skupine istraživanja generiran je iz obrisanih podataka koristeći funkciju raspršivanja. Iako su ime, adresa i datum rođenja izbrisani iz tablice, ako su ime, adresa i datum rođenja osobe čiji se podaci obrađuju poznati uz poznavanje korištene funkcije raspršivanja, jednostavno je izračunati referentne brojeve skupine istraživanja.

- *Društvene mreže*

Prikazano je[19] da se osjetljivi podaci o određenim pojedincima mogu izvesti iz grafikona društvenih mreža, unatoč tehnikama „pseudonimizacije" primijenjenim na takve podatke. Pružatelj društvene mreže netočno je pretpostavio da je pseudonimizacija bila pouzdana da spriječi utvrđivanje identiteta nakon prodaje podataka drugim trgovačkim društvima u marketinške i reklamne svrhe. Umjesto stvarnih imena pružatelj je koristio nadimke, ali to očito nije bilo dovoljno za anonimizaciju korisničkih profila, jer su odnosi između različitih pojedinaca jedinstveni i mogu se koristiti kao identifikator.

- *Lokacije*

Znanstvenici na MIT-u[20] nedavno su analizirali pseudonimizirani skup podataka koji se sastoji od 15 mjeseci prostorno-vremenskih koordinata kretanja 1,5 milijuna ljudi na području s radijusom od 100 km. Pokazali su da je sa četiri lokacijske točke bilo moguće izdvojiti 95 % populacije, a da su bile dovoljne samo dvije točke za izdvajanje više od 50 % osoba čiji se podaci obrađuju (jedna od takvih točaka je poznata, vrlo vjerojatno je to „kuća" ili „ured") s vrlo ograničenim prostorom za zaštitu privatnosti, čak i da su identiteti pojedinaca pseudonimizirani zamjenom njihovih pravih atributa [….] drugim oznakama.

5. Zaključci i preporuke

5.1. Zaključci

Tehnike deidentifikacije i anonimizacije predmetom su intenzivnog istraživanja, a u ovom je dokumentu dosljedno prikazano da svaka tehnika ima svojih prednosti i nedostataka. U većini slučajeva nije moguće dati minimalne preporuke za parametre koji će se koristiti jer u svakom skupu podataka treba razmatrati od slučaja do slučaja.

U mnogim slučajevima anonimizirani skup podataka i dalje može predstavljati rezidualni rizik za osobe čiji se podaci obrađuju. I zaista, čak i kada više nije moguće precizno dohvatiti zapis pojedinca, postoji mogućnost skupljanja podataka o tom pojedincu pomoću drugih dostupnih izvora podataka (javno ili ne). Treba naglasiti da se osim izravnog učinka na osobe čiji se podaci obrađuju koji je prouzročen lošim procesom anonimizacije (dosađivanje, trošak vremena i osjećaj gubitka kontrole zbog uključenosti u klaster bez znanja ili prethodnog pristanka), ostale neizravne popratne pojave loše anonimizacije mogu pojaviti kadgod je neki napadač pogrešno uključio u metu osobu čiji se podaci obrađuju, kao posljedica obrade anonimiziranih podataka – posebno ako su namjere napadača zlonamjerne. Stoga radna skupina naglašava da tehnike anonimizacije mogu pružiti jamstva privatnosti, ali samo ako se njihova primjena organizira na odgovarajući način – što znači da jasno treba odrediti preduvjete (kontekst) i cilj/ciljeve procesa anonimizacije kako bi se ostvarila ciljana razina anonimizacije.

[19] A. Narayanan i V. Šmatikov, „Deanonimizacija društvenih mreža", na 30. Simpoziju IEEE-a o sigurnosti i privatnosti, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen i V. Blondel, „Jedinstven u gomili: Ograničenja privatnosti ljudske mobilnosti", Nature, br. 1376, 2013.

5.2. Preporuke

- Neke tehnike anonimizacije pokazuju svojstvena ograničenja. Ta ograničenja treba ozbiljno razmotriti prije nego što nadzornici podataka koriste danu tehniku za oblikovanje procesa anonimizacije. Moraju uzeti u obzir koje svrhe treba postići anonimizacijom – poput zaštite privatnosti pojedinaca kod objave skupa podataka, ili dopuštanja dohvaćanja dijela podatke iz skupa podataka.
- Tehnike opisane u ovom dokumentu ne ispunjuju sa sigurnošću kriterije učinkovite anonimizacije (tj. bez izdvajanja pojedinca; bez povezivosti između zapisa povezanih s pojedincem; i bez izvođenja zaključaka o pojedincu). Međutim, neke od ovih rizika moguće je u cijelosti ili dijelom izbjeći nekom od tehnika, potrebno je posebno planiranje prilikom osmišljavanja primjene pojedinačne tehnike za posebnu situaciju i primjene kombinacije tih tehnika kao način povećanja pouzdanosti rezultata.

U sljedećoj je tablici prikazan pregled prednosti i nedostataka razmatranih tehnika u pogledu tri osnovna uvjeta:

	Je li izdvajanje i dalje rizik?	**Je li povezivost i dalje rizik?**	**Je li izvođenje zaključaka i dalje rizik?**
Pseudonimizacija	Da	Da	Da
Dodavanje šuma	Da	Ne mora biti	Ne mora biti
Zamjena	Da	Da	Ne mora biti
Sažimanje ili K-anonimnost	Ne	Da	Da
L-raznolikost	Ne	Da	Ne mora biti
Diferencijalna privatnost	Ne mora biti	Ne mora biti	Ne mora biti
Raspršivanje/tokenizacija	Da	Da	Ne mora biti

Tablica 6. Prednosti i mane razmatranih tehnika

- O optimalnom rješenju treba odlučiti na temelju pojedinačnog slučaja. Rješenje (tj. proces potpune anonimizacije) kojim se zadovoljuju tri kriterija bilo bi pouzdano u pogledu utvrđivanja identiteta provedenog najvjerojatnijim i najopravdanijim sredstvima koje nadzornik podataka ili bilo koja treća stranka mogu koristiti.
- Kada prijedlogom nije zadovoljen jedan od kriterija, treba provesti detaljnu evaluaciju rizika utvrđivanja identiteta. Tu bi evaluaciju trebalo podnijeti nadležnom tijelu ako, u skladu s nadležnim zakonodavstvom, nadležno tijelo mora procijeniti ili odobriti proces anonimizacije.

Za smanjenje rizika utvrđivanja identiteta treba razmotriti sljedeće dobre prakse:

<u>Dobre prakse anonimizacije</u>

Općenito:

- Ne treba se pouzdati u pristup „objavi i zaboravi”. S obzirom na rezidualni rizik utvrđivanja identiteta, nadzornici podataka bi trebali:
    - 1. utvrđivati nove rizike i redovito ponovno evaluirati rezidualni/rezidualne rizik/rizike,
    - 2. procijeniti jesu li kontrole za utvrđene rizike dovoljne i prilagoditi ih u skladu s time i

- 3. pratiti i kontrolirati rizike.

- Kao dio tih rezidualnih rizika, uzimati u obzir potencijal utvrđivanja neanonimiziranog dijela skupa podataka (ako postoji), posebno u kombinaciji s anonimiziranim dijelom, i potencijal mogućih korelacija između atributa (npr. između geografske lokacije i podataka o razini bogatstva).

Kontekstni elementi:

- Svrhe koje treba ostvariti anonimiziranim skupom podataka treba jasno utvrditi jer imaju ključnu ulogu u utvrđivanju rizika utvrđivanja identiteta.
- Ovo ide ruku pod ruku s razmatranjem svih mjerodavnih kontekstnih elemenata – npr. prirode originalnih podataka, kontrolnih mehanizama koji se koriste (uključujući sigurnosne mjere za ograničavanje pristupa skupovima podataka), veličine uzorka (kvantitativna svojstva), dostupnosti sredstava javnog informiranja (u koja se primatelji mogu pouzdati), predviđene objave podataka trećim strankama (ograničene, neograničene npr. na internetu itd.).
- Treba posvetiti pažnju mogućim napadačima uzimajući u obzir privlačnost podataka za ciljane napade (ponovno, osjetljivost podataka i prirode podataka bit će ključni čimbenici s time u vezi).

Tehnički elementi:

- Nadzornici podataka trebali bi objaviti tehniku anonimizacije / mješavinu tehnika koje se provode, posebno ako planiraju objaviti anonimizirani skup podataka.
- Očite (npr. rijetke) atribute / kvazi-identifikatore trebalo bi ukloniti iz skupa podataka.
- Ako se koriste tehnike dodavanja šuma (u randomizaciji), razinu šuma dodanu zapisima treba utvrditi kao funkciju vrijednosti atributa (dakle ne treba ubacivati disproporcionalni šum), treba zaštititi utjecaj na osobe čiji se podaci obrađuju, i/ili raštrkanost skupa podataka.
- Kod oslanjanja na diferencijalnu privatnost (u randomizaciji), treba uzeti u obzir potrebu praćenja upita kako bi se otkrili upiti koji ometaju privatnost jer je ometajuće svojstvo upita kumulativno.
- Ako se provode tehnike generalizacije, bitno je da se nadzornici podataka ne ograničuju na jedan kriterij generalizacije čak i za isti atribut; to znači da treba odabrati različite zrnatosti lokacije ili različite vremenske intervale. Odabir kriterija koji će se primjenjivati mora ovisiti o distribuciji vrijednosti atributa u danoj populaciji. Ne mogu se sve distribucije generalizirati – tj. u generalizaciji se ne može koristiti isti pristup u svim slučajevima. Treba osigurati varijabilnost unutar klase ekvivalencije; na primjer, treba odabrati određeni prag ovisno o „kontekstnim elementima" spomenutima iznad (veličina uzorka itd.) i ako se ne dostigne taj prag, tada određeni uzorak treba odbaciti (ili treba postaviti drugačiji kriterij generalizacije).

PRILOG
Priručnik o tehnikama anonimizacije

A.1. Uvod

Anonimnost se različito tumači širom EU-a – pri čemu u nekim zemljama predstavlja računsku anonimnost (tj. trebalo bi biti računski teško, čak i nadzorniku u suradnji s bilo kojom strankom, izravno ili neizravno utvrditi identitet jedne od osoba čiji se podaci obrađuju), a u drugim zemljama savršenu anonimnost (tj. trebalo bi biti nemoguće, čak i nadzorniku u suradnji s bilo kojom strankom, izravno ili neizravno utvrditi identitet jedne od osoba čiji se podaci obrađuju). Unatoč tomu, „anonimizacija" u oba slučaja predstavlja proces kojim se podaci čine anonimnima. Razlika je u tome što se smatra prihvatljivim rizikom ponovnog utvrđivanja identiteta.

Anonimizirani podaci mogu imati raznovrsnu primjenu, od društvenih istraživanja, statističkih analiza, razvoja novih usluga/proizvoda. Ponekad čak i aktivnosti opće svrhe mogu imati utjecaj na određene osobe čiji se podaci obrađuju, poništavajući navodnu anonimnu prirodu obrađenih podataka. Moguće je dati mnogo primjera, od pokretanja ciljanih marketinških inicijativa do provedbe javnih mjera koje se temelje na izradi profila, ili ponašanja, ili obrazaca kretanja[21].

Nažalost, osim općih izjava ne postoji potpuno razvijena metrika za ocjenjivanje unaprijed potrebnog vremena ili truda za ponovno utvrđivanje identiteta nakon obrade, ili alternativno odabrati najprikladniji postupak za uspostavljanje ako se želi smanjiti potrebu da se objavljena baza podataka odnosi na utvrđeni skup osoba čiji se podaci obrađuju.

„Umjetnost anonimizacije", kako se ove prakse ponekad nazivaju u znanstvenoj literaturi[22], nova je znanstvena grana koja je još u povojima i postoje mnoge prakse za smanjenje moći utvrđivanja skupova podataka. Međutim, treba istaknuti da se većinom takvih praksi ne sprečava povezivanje obrađenih podataka s osobama čiji se podaci obrađuju. U nekim se okolnostima utvrđivanje skupova podataka koji se smatraju anonimnima dokazalo vrlo uspješnim, u ostalim situacijama pojavili su se lažni pozitivni rezultati.

Općenito govoreći, postoje dva različita pristupa: jedan se temelji na generalizaciji atributa, a drugi na randomizaciji. Analizom pojedinosti i razrađenosti tih praksi doći ćemo do novog uvida o moći utvrđivanja podataka i prikazati sam pojam osobnih podataka u novom svjetlu.

A.2. „Anonimizacija" pomoću randomizacije

Jedna mogućnost anonimizacije sastoji se od izmjene trenutačnih vrijednosti za sprečavanje povezivanja između anonimiziranih podataka i originalnih vrijednosti. Ovaj se cilj može postići velikim brojem metoda od ubrizgavanja šuma do zamjene podataka (permutacije). Treba naglasiti da je uklanjanje atributa istovjetno ekstremnom obliku randomizacije tog atributa (koji je atribut u potpunosti pokriven šumom).

U nekim okolnostima cilj cjelokupne obrade nije objava randomiziranog skupa podataka, već dopuštanje pristupa podacima pomoću upita. Rizik za osobu čiji se podaci obrađuju u ovom slučaju dolazi iz vjerojatnosti da će napadač imati mogućnost izvući podatke iz niza različitih

[21] Na primjer slučaj TomTom u Nizozemskoj (vidi primjer objašnjen u stavku 2.2.3).

[22] Jun Gu, Yuexian Chen, Junning Fu, Huanchun Peng, Xiaojun Ye, Sintetiziranje: Umjetnost anonimizacije, bilješke s predavanja o bazi podataka i aplikacijama stručnih sustava u računalnoj znanosti –Springer- Svezak 6261, 2010., str 385. – 399.

upita bez znanja nadzornika podataka. Kako bi se pojedincima u skupu podataka zajamčila anonimnost, ne bi trebalo biti moguće zaključiti da je osoba čiji se podaci obrađuju doprinijela skupu podataka, te tako prekidajući vezu s bilo kojom vrstom prethodnih podataka koje napadač može imati.

Po potrebi se dodavanjem šuma odgovoru na upit može dodatno smanjiti rizik od ponovnog utvrđivanja identiteta. Ovaj pristup, također poznat u literaturi kao diferencijalna privatnost[23], polazi od onoga opisanog ranije jer se osobama koje objavljuju podatke daje veća kontrola nad pristupom podacima u usporedbi s javnom objavom. Dodavanje šuma ima dva glavna cilja: jedan je zaštita privatnosti osoba čiji se podaci obrađuju u skupu podataka, a drugi zadržavanje korisnosti objavljenih podataka. Jačina šuma mora posebno biti proporcionalna razini postavljanja upita (previše upita o pojedincima na koje treba odgovoriti previše precizno rezultira povećanjem vjerojatnosti utvrđivanja identiteta). Uspješnu primjenu randomizacije danas treba razmatrati od slučaja do slučaja bez tehnike koja nudi jednostavnu metodologiju jer postoje primjeri curenja podataka o atributima osobe čiji se podaci obrađuju (bilo da je ovo uključeno u skup podataka ili ne) čak i kada se smatralo da je nadzornik podataka randomizirao skup podataka.

Može biti korisno razmotriti specifične primjere za objašnjenje potencijalnih propusta randomizacije kao sredstava za omogućivanje anonimizacije. Na primjer, u kontekstu interaktivnog pristupa, upiti kojima se štiti privatnost mogu predstavljati rizik za osobe čiji se podaci obrađuju. Zapravo, ako napadač zna da je podskupina S pojedinaca unutar skupa podataka koji sadržava informacije o učestalosti atributa A u populaciji P, jednostavnim upitom s dva pitanja „Koliko pojedinaca u populaciji P ima atribut A?" i „Koliko pojedinaca u populaciji P, osim onih koji pripadaju podskupini S, ima atribut A?" postoji mogućnost utvrđivanja (razlikom) broja pojedinaca u S koji zaista imaju atribut A – ili deterministički ili pomoću vjerojatnosti učestalosti. U svakom slučaju, privatnost pojedinaca u podskupini S može biti ozbiljno ugrožena, posebno ovisno o prirodi atributa A.

Također se može razmotriti da ako osoba čiji se podaci obrađuju nije u skupu podataka, već je poznat njezin odnos s podacima unutar skupa podataka, tada objava skupa podataka može izazvati rizik za njezinu sigurnost. Na primjer, ako je poznato da se „ciljna vrijednost atributa A razlikuje količinom X od prosječne vrijednosti populacije", ako napadač jednostavno zatraži od osobe koja upravlja bazom podataka da obavi postupak izvođenja prosječne vrijednosti atributa A kojim se štiti privatnost, može točno izvesti zaključak o osobnim podacima koji se odnose na određenu osobu čiji se podaci obrađuju.

Ubacivanje određene količine relativnih netočnosti u aktualne vrijednosti u bazu podataka jest operacija koju treba ispravno oblikovati. Treba dodati dovoljno šuma kako bi se zaštitila privatnost, ali također dovoljno malo da se očuva korisnost podataka. Na primjer, ako je broj osoba čiji se podaci obrađuju s čudnim atributom vrlo mali ili s visokom osjetljivosti atributa, može biti bolje prijaviti raspon ili generičku rečenicu poput „mali broj slučajeva, moguće čak i nula", umjesto prijavljivanja aktualnog broja. Na taj je način, čak i ako je bučni mehanizam objave poznat unaprijed, očuvana privatnost osobe čiji se podaci obrađuju, jer ostaje stupanj nesigurnosti. Iz perspektive korisnosti, ako se netočnost ispravno oblikuje, rezultati su i dalje korisni za statističke svrhe ili donošenje odluka.

Nadalje treba razmotriti randomizaciju baze podataka i pristup diferencijalne privatnosti. Prvo, prava količina iskrivljenja može se značajno razlikovati u odnosu na kontekst (vrsta

[23] Cynthia Dwork, Diferencijalna privatnost, Međunarodni kolokvij o automatima, jezicima i programiranju (ICALP) 2006., str. 1. – 12.

upita, veličina populacije u bazi podataka, priroda atributa i njegova svojstvena snaga utvrđivanja) i ne može se predvidjeti rješenje"*ad omnia*". Nadalje, kontekst se s vremenom može mijenjati i s time u skladu trebalo bi izmijeniti interaktivni mehanizam. Podešavanje šuma zahtijeva praćenje kumulativnih rizika privatnosti koje svaki interaktivni mehanizam predstavlja za osobe čiji se podaci obrađuju. Mehanizam pristupa podacima tada bi trebalo opremiti upozorenjima kada se dosegne proračun „troška privatnosti" i osobe čiji se podaci objavljuju mogle bi biti izložene posebnim rizicima ako se novi upit nastavi, kako bi se nadzorniku podataka pomoglo pri utvrđivanju odgovarajuće razine iskrivljenja koje svaki put treba ubaciti u aktualne osobne podatke.

S druge strane, također bismo trebali razmotriti slučaj kada su vrijednosti atributa izbrisane (ili izmijenjene). Rješenje koje se obično koristi za neke atipične vrijednosti za atribute jest brisanje skupa podataka koji se odnosi na atipične pojedince ili brisanje atipičnih vrijednosti. U posljednjem slučaju važno je osigurati da samo nepostojanje vrijednosti ne postaje element za utvrđivanje identiteta osobe čiji se podaci obrađuju.

Razmotrimo sada randomizaciju pomoću zamjene atributa. Najveća pogrešna predodžba kod anonimizacije jest njezino izjednačavanje s enkripcijom ili kodiranjem ključa. Ova se pogrešna predodžba temelji na dvije pretpostavke – naime a) da kada se enkripcija primijeni na neke atribute zapisa u bazi podataka (npr. ime, adresu, datum rođenja) ili se ti atributi zamijene naizgled randomiziranim nizom kao rezultat postupka kodiranja ključa poput funkcije raspršivanja ključem, tada je taj zapis „anonimiziran" i b) da je anonimizacija učinkovitija ako je duljina ključa odgovarajuća i enkripcijski algoritam je najmoderniji. Ova pogrešna predodžba je jako raširena među nadzornicima podataka i zaslužuje objašnjenje, jer je također povezana s pseudonimizacijom i njezinim navodnim manjim rizicima.

Prvo, ciljevi ovih tehnika radikalno su različiti: cilj je enkripcije kao sigurnosne prakse omogućiti povjerljivost komunikacijskog kanala između utvrđenih stranaka (ljudska bića, uređaji, ili dijelovi softvera/hardvera) kako bi se izbjeglo prisluškivanje ili nenamjerno objavljivanje. Kodiranje ključa odgovara semantičkoj translaciji podataka ovisno o tajnom ključu. S druge strane, cilj je anonimizacije izbjeći utvrđivanje identiteta pojedinaca sprečavanjem skrivenog povezivanja atributa s osobom čiji se podaci obrađuju.

Ni samom enkripcijom ni kodiranjem ključa ne može se učiniti neutvrdivim identitet osobe čiji se podaci obrađuju: jer su, barem u rukama nadzornika, originalni podaci i dalje dostupni ili se o njima može izvesti zaključak. Jedina provedba semantičke translacije osobnih podataka, kako se događa s kodiranjem ključa, ne uklanja mogućnost vraćanja podataka u njihovu originalnu strukturu – primjenom algoritma u suprotnom smjeru ili brutalnim napadima, ovisno o prirodi shema, ili kao rezultat povrede podataka. Najmodernijom enkripcijom može se osigurati da su podaci zaštićeni do višeg stupnja, tj. nerazumljivi su za subjekte koji ne znaju ključ enkripcije, ali ne rezultira nužno anonimizacijom. Dok god su dostupni ključ ili originalni podaci (čak i u slučaju povjerljive treće stranke, obvezane ugovorom na pružanje usluge sigurnog pohranjivanja ključa), još uvijek postoji mogućnost utvrđivanja identiteta osobe čiji se podaci obrađuju.

Usmjerenost samo na pouzdanost enkripcijskog mehanizma kao mjere stupnja „anonimizacije" skupa podataka zavaravajuće je jer na cjelokupnu sigurnost enkripcijskog mehanizma ili funkcije raspršivanja utječu mnogi drugi tehnički i organizacijski čimbenici . U literaturi su navedeni mnogi uspješni napadi koji u potpunosti zaobilaze algoritam, ili zato što utječu na slabost u nadzoru ključeva (npr. postojanje nesigurnijeg zadanog načina) ili zbog ostalih ljudskih faktora (npr. slabe lozinke za vraćanje ključa). Konačno, odabrana enkripcijska shema s danom veličinom ključa osmišljenja je u cilju osiguranja povjerljivosti

za dano razdoblje (većini aktualnih ključeva trebat će promijeniti veličinu oko 2020.), dok proces anonimizacije ne bi trebao biti vremenski ograničen.

Sada treba razraditi granice randomizacije atributa (ili zamjene i uklanjanja) uzimajući u obzir razne loše primjere anonimizacije randomizacijom koji su se pojavili zadnjih godina i razloge tih neuspjeha.

Dobro poznat slučaj o objavi loše anonimiziranog skupa podataka jest onaj o nagradi Netflix[24]. Gledajući generički zapis u bazi podataka u kojoj je randomiziran određen broj atributa koji se odnose na osobu čiji se podaci obrađuju, svaki se zapis i dalje može podijeliti na dva podzapisa kako slijedi: {randomizirani atributi, čisti atributi}, pri čemu jasni atributi mogu biti bilo koja kombinacija navodno neosobnih podataka. Posebno zapažanje koje se može vidjeti iz skupa podataka nagrade Netflix dolazi iz razmatranja da svaki zapis može biti predstavljen točkom u višedimenzionalnom prostoru, pri čemu je svaki čisti atribut koordinata. Korištenjem ove tehnike svaki se skup podataka može smatrati konstelacijom točaka u takvom višedimenzionalnom prostoru koji može pokazati visoki stupanj raštrkanosti, što znači da točke mogu biti udaljene jedna od druge. One zaista mogu biti toliko udaljene da nakon podjele prostora u široka područja svaka regija sadržava samo jedan zapis. Čak i ubacivanje šuma ne uspijeva dovoljno približiti zapise kako bi dijelili to isto višedimenzionalno područje. Na primjer, u eksperimentu Netflix zapisi su bili dovoljno jedinstveni sa samo 8 ocjena filmova danih unutar 14 dana razmaka. Nakon dodavanja šuma i ocjenama i datumima, nije bilo moguće primijetiti preklapanje područja. Drugim riječima, potpuno isti odabir od samo 8 ocijenjenih filmova predstavljao je otisak prsta izraženih ocjena, koji ne dijele dvije osobe čiji se podaci obrađuju u bazi podataka. Na temelju ovog geometrijskog zapažanja istraživači su uskladili navodno anoniman skup podataka Netflix s drugom javnom bazom podataka s ocjenama filmova (IMDB), i pronašli korisnike koji su ocijenili iste filmove u istim vremenskim intervalima. Budući da je većina korisnika pokazala podudarnost jedan prema jedan, bilo je moguće importirati pomoćne podatke dohvaćene iz baze podataka IMDB u objavljeni skup podataka Netflix te tako obogatiti identitetima sve navodno anonimizirane zapise.

Važno je naglasiti sljedeće opće svojstvo: rezidualni dio bilo koje „randomizirane" baze podataka i dalje posjeduje vrlo veliku moć utvrđivanja identiteta, ovisno o rijetkosti kombinacije rezidualnih atributa. Ovo je upozorenje koje nadzornici podataka uvijek trebaju imati na umu prilikom odabira randomizacije kao njihova načina postizanja ciljane anonimizacije.

Mnogi eksperimenti ponovnog utvrđivanja identiteta ove vrste slijedili su sličan pristup projekcije dviju baza podataka u isti podprostor. Ovo je vrlo snažna metodologija ponovnog utvrđivanja identiteta, koja je nedavno imala mnogo primjena u različitim područjima. Na primjer, eksperiment utvrđivanja identiteta proveden u društvenoj mreži[25] iskoristio je društveni grafikon korisnika pseudonimiziranih pomoću oznaka. U ovom slučaju, atributi korišteni za utvrđivanje identiteta bili su popis kontakata svakog korisnika jer je prikazano da je vrlo mala vjerojatnost identičnog popisa kontakata između dva pojedinca. Na temelju ove intuitivne pretpostavke utvrđeno je da podgrafikon unutarnjih veza vrlo ograničenog broja čvorova predstavlja topološki otisak prsta za vraćanje, skriven unutar mreže, i da se veliki dio

[24] Arvind Narayanan, Vitalij Šmatikov: Pouzdana deanonimizacija velikih raštrkanih skupova podataka. Simpozij IEEE-a o sigurnosti i privatnosti 2008: str. 111. – 125.
[25] L. Backstrom, C. Dwork i J. M. Kleinberg. *Wherefore art thou r3579x?*: anonimizirane društvene mreže, skriveni obrasci i strukturalna steganografija, zapisnik sa 16. međunarodne konferencije o World Wide Webu WWW'07, str. 181. – 190. (2007.)

cijele društvene mreže može utvrditi kada se utvrdi ova podmreža. Slijede podatke o uspjesima sličnog napada: prikazano je da korištenjem manje od 10 čvorova (koji mogu prouzročiti milijun različitih konfiguracija podmreže, pri čemu svaka potencijalno predstavlja topološki otisak prsta) društvena mreža s više od 4 milijuna pseudonimiziranih čvorova i 70 milijuna poveznica može biti sklona napadima ponovnog utvrđivanja identiteta, te se može ugroziti privatnost velikog broja veza. Treba naglasiti da ovaj pristup ponovnog utvrđivanja identiteta nije dizajniran po mjeri posebnog konteksta društvenih mreža, ali je dovoljno općenit da se potencijalno može prilagoditi ostalim bazama podataka u kojima se bilježe odnosi između korisnika (npr. telefonski kontakti, korespondencije e-pošte, internetske stranice za pronalaženje partnera itd.).

Drugi način utvrđivanja navodno anonimnog zapisa temelji se na analizi stila pisanja (stilometrija)[26]. Već je razvijen određeni broj algoritama kojim se izvlači metrika iz raščlanjenog teksta uključujući učestalost korištenja određenih riječi, pojavljivanje posebnih gramatičkih obrazaca i vrste interpunkcije. Sva se ova svojstva mogu koristiti za povezivanje navodno anonimnog teksta sa stilom pisanja utvrđenog autora. Istraživači su izvukli stil pisanja s više od 100 000 blogova i danas mogu automatski utvrditi autora objave s preciznošću od gotovo 80 %; očekuje se da će preciznost ove tehnike dalje rasti i istraživati ostale signale poput lokacije ili ostalih metapodataka koji se nalaze u tekstu.

Moć utvrđivanja identiteta korištenjem semantike zapisa (tj. rezidualnog nerandomiziranog dijela zapisa) pitanje je koje zaslužuje više razmatranja od strane istraživačke zajednice i industrije. Nedavno vraćanje identiteta darovatelja DNK (2013.)[27] pokazuje da je ostvaren vrlo mali napredak od poznatog incidenta s AOL-om (2006.) – kada je javno objavljena baza podataka s dvadeset milijuna ključnih riječi pretraživanja za više od 650 000 korisnika u tromjesečnom razdoblju. To je rezultiralo utvrđivanjem identiteta i lokacije određenog broja korisnika AOL-a.

Druga obitelj podataka koji se rijetko anonimiziraju samo uklanjanjem identiteta osoba čiji se podaci obrađuju ili djelomičnom enkripcijom nekih atributa su podaci o lokaciji. Obrasci kretanja ljudskih bića mogli bi biti dovoljno jedinstveni da se na temelju semantičkog dijela podataka o lokaciji (mjestima gdje se osoba nalazila u određeno vrijeme), čak i bez ostalih atributa, mogu otkriti mnoge osobine osobe čiji se podaci obrađuju[28]. Ovo je više puta dokazano u reprezentativnim akademskim studijama[29].

S tim u vezi potrebno je upozoriti na korištenje pseudonima kao načina pružanja odgovarajuće zaštite osobama čiji se podaci obrađuju od curenja identiteta ili atributa. Ako se pseudonimizacija temelji na zamjeni identiteta drugim jedinstvenim kodom, pretpostavka da ovo predstavlja pouzdanu deidentifikaciju je naivna i ne uzima u obzir složenost metodologija utvrđivanja identiteta i raznovrsne kontekste gdje bi se mogle primijeniti.

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

[27] Genetski podaci su posebno značajan primjer osjetljivih podataka čija reidentifikacija može biti ugrožena ako je jedini mehanizam za njihovu „anonimizaciju“ uklanjanje identiteta darivatelja. Vidjeti primjer citiran u stavku 2.2.2. Vidjeti također John Bohannon, Genealoške baze podataka omogućuju imenovanje anonimnih donatora DNK, Science, Svezak. 339, br. 6117 (18. siječnja 2013.), str. 262.

[28] Ovo je pitanje obrađivano u nekim nacionalnim zakonodavstvima. Na primjer, u Francuskoj se objavljeni statistički podaci o lokaciji anonimiziraju pomoću generalizacije i permutacije. Prema tome, INSEE objavljuje statističke podatke koji su generalizirani sažimanjem svih podataka na područje od 40 000 kvadratnih metara. Zrnatost skupa podataka je dovoljna da očuva korisnost podataka, a permutacije sprečavaju napade deanonimizacije u raštrkanim područjima. Općenitije, sažimanje ove obitelji podataka i njihova permutacija pruža snažna jamstva od napada izvođenja zaključaka i deanonimizacije (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. & Blondel, V.D. Jedinstven u gomili: Ograničenja privatnosti ljudske mobilnosti. Nature. 3, 1376 (2013.).

A.3. „Anonimizacija” pomoću generalizacije

Jednostavnim se primjerom može pomoći u objašnjenju pristupa koji se temelji na generalizaciji atributa.

Razmotrimo slučaj u kojem nadzornik podataka odluči objaviti jednostavnu tablicu koja sadržava tri informacije ili atributa: identifikacijski broj jedinstven za svaki zapis, identifikator lokacije koji povezuje osobu čiji se podaci obrađuju s mjestom gdje živi i identifikator svojstva koji prikazuje svojstvo koje ima osoba čiji se podaci obrađuju; pretpostavimo još da je ovo svojstvo jedna od dvije različite vrijednosti, obično navedene kao {P1, P2} :

Serijski broj	**Identifikator lokacije**	**Svojstvo**
#1	Rim	P1
#2	Madrid	P1
#3	London	P2
#4	Pariz	P1
#5	Barcelona	P1
#6	Milano	P2
#7	New York	P2
#8	Berlin	P1

Tablica A1. Uzorak osoba čiji se podaci obrađuju prikupljen po lokaciji i svojstvima P1 i P2

Ako netko koga nazivamo napadač unaprijed zna da je određena osoba čiji se podaci obrađuju (subjekt) koja živi u Milanu uključena u tablicu, tada nakon pregleda tablice on može saznati da također posjeduje svojstvo P2, budući da je #6 jedina osoba s tim identifikatorom lokacije.

Ovim su vrlo jednostavnim primjerom prikazani glavni elementi svakog postupka utvrđivanja identiteta koji se primjenjuje na skup podataka koji je prošao kroz proces navodne anonimizacije. Naime, postoji napadač koji je (slučajno ili namjerno) imao prethodno znanje o nekim ili svim osobama čiji se podaci obrađuju u skupu podataka. Napadač želi povezati prethodno znanje s podacima u objavljenom skupu podataka kako bi dobio jasniju sliku obilježja tih osoba čiji se podaci obrađuju.

Kako bi se smanjila učinkovitost ili neodgodivost povezivanja podataka s bilo kojim pozadinskim znanjem, nadzornik podataka mogao bi se usredotočiti na identifikator lokacije tako da zamijeni grad u kojem žive osobe čiji se podaci obrađuju sa širim područjem poput zemlje. Na taj bi način tablica izgledala kako slijedi.

Serijski broj	**Identifikator lokacije**	**Svojstvo**
#1	Italija	P1
#2	Španjolska	P1
#3	UK	P2
#4	Francuska	P1
#5	Španjolska	P1
#6	Italija	P2
#7	SAD	P2
#8	Njemačka	P1

Tablica A2. Generalizacija tablice A1 po nacionalnosti

S ovim novim sažimanjem podataka prethodno znanje napadača o utvrđenoj osobi čiji se podaci obrađuju (npr. „subjekt živi u Rimu i nalazi se u tablici") ne omogućuje se izvođenje nikakvog jasnog zaključka o njegovu svojstvu: to je zato što dva Talijana u tablici imaju različita svojstva, P1 odnosno P2. Napadač ima 50 % nesigurnosti o svojstvu ciljnog subjekta. Ovaj jednostavni primjer pokazuje učinak generalizacije na praksu anonimizacije. Zapravo, iako ovo lukavstvo generalizacije može biti učinkovito da prepolovi vjerojatnost utvrđivanja talijanskog subjekta, nije učinkovit za subjekt iz drugih mjesta (npr. SAD-a).

Nadalje, napadač i dalje može saznati podatke o španjolskom subjektu. Ako je prethodno znanje tipa „subjekt živi u Madridu i nalazi se u tablici" ili „subjekt živi u Barceloni i nalazi se u tablici", napadač može sa 100 %-tnom sigurnošću zaključiti da cilj posjeduje svojstvo P1. Stoga generalizacija ne donosi istu razinu privatnosti ili otpornosti na napade izvođenja zaključaka za cijelu populaciju u skupu podataka.

Ako slijedimo ovu argumentaciju, možemo biti u iskušenju zaključiti da bi snažnija generalizacija mogla pomoći u sprečavanju povezivanja – na primjer generalizacija po kontinentu. Na taj bi način tablica izgledala ovako:

Serijski broj	**Identifikator lokacije**	**Svojstvo**
#1	Europa	P1
#2	Europa	P1
#3	Europa	P2
#4	Europa	P1
#5	Europa	P1
#6	Europa	P2
#7	Sjeverna Amerika	P2
#8	Europa	P1

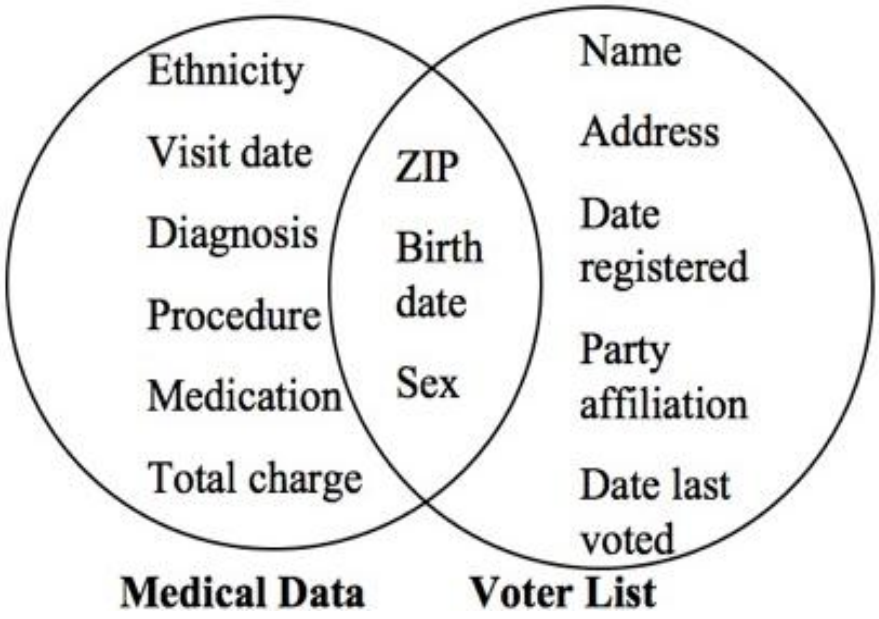
Tablica A3. Generalizacija tablice A1 po kontinentu

S ovom vrstom sažimanja sve osobe čiji se podaci obrađuju u tablici, osim one koja živi u SAD-u, bile bi zaštićene od napada povezivanja i utvrđivanja identiteta te bi sve prethodne informacije tipa „subjekt živi u Madridu i nalazi se u tablici" ili „subjekt živi u Milanu i nalazi se u tablici" dovele do neke razine vjerojatnosti kao za svojstvo koje se primjenjuje na danu osobu čiji se podaci obrađuju (P1 s vjerojatnošću od 71,4 % i P2 s vjerojatnošću od 28,6 %), nego do izravnog povezivanja. Također, ovakvom daljnjom generalizacijom dolazi do očitog i radikalnog gubitka podataka: tablica ne dopušta otkrivanje potencijalnih korelacija između svojstava i lokacije, naime može li određena lokacija s većom vjerojatnošću aktivirati bilo koje od dva svojstva, jer dopušta samo takozvane „marginalne" distribucije, naime apsolutnu vjerojatnost pojavljivanja svojstva P1 i P2 širom populacije (62,5 % odnosno 37,5 % u našem primjeru) i unutar svakog kontinenta (kao što je rečeno 71,4 % i 28,6 % u Europi i 100 % i 0 % u Sjevernoj Americi.

Iz primjera se također može zaključiti da praksa generalizacije utječe na praktičnu korisnost podataka. Neki inženjerski alati dostupni su od danas za suočavanje unaprijed (tj. prije objave skupa podataka) s onim što je najprikladnija razina generalizacije atributa, kako bi se smanjili rizici utvrđivanja identiteta za osobe čiji se podaci obrađuju unutar tablice bez utjecaja na korisnost objavljenih podataka u prevelikoj mjeri.

k-anonimnost

Pokušaj sprečavanja napada povezivanja, na temelju generalizacije atributa, poznat je kao k-anonimnost. Ta praksa potječe iz eksperimenta ponovnog utvrđivanja identiteta provedenog krajem 1990.-ih, kada je privatno američko trgovačko društvo iz zdravstvenog sektora javno objavilo navodno anonimizirani skup podataka. Ta se anonimizacija sastojala od uklanjanja imena osoba čiji se podaci obrađuju, ali skup podataka je i dalje sadržavao zdravstvene podatke i ostale atribute poput poštanskog broja (identifikator lokacije gdje su živjele), spol i potpuni datum rođenja. Ista skupina od tri informacije {poštanski broj, spol, potpuni datum rođenja} također je bila uključena u ostalim javno dostupnim evidencijama (npr. popis birača) i stoga neki akademski znanstvenik mogao je koristiti za povezivanje identiteta određenih osoba čiji se podaci obrađuju s atributima u objavljenom skupu podataka. Prethodno znanje koje je imao napadač (istraživač) moglo bi biti sljedeće: „Znam da je osoba čiji se podaci obrađuju s popisa birača s određenom skupinom od tri informacije {poštanski broj, spol, potpuni datum rođenja} jedinstvena. Postoji zapis u objavljenom skupu podataka s tom skupinom od tri informacije". Empirijski je primijećeno[30] da je velika većina (više od 80 %) osoba čiji se podaci obrađuju u javnoj evidenciji korištenoj u ovom istraživačkom eksperimentu jednoznačno povezana s određenom skupinom od tri podatka, što je omogućilo utvrđivanje identiteta. U skladu s time, podaci u ovom slučaju nisu ispravno anonimizirani.



Slika A1. Ponovno utvrđivanje identiteta povezivanjem podataka

Kako bi se smanjila učinkovitost sličnih napada povezivanja, rečeno je da bi nadzornici prvo trebali pregledati skup podataka i grupirati one atribute koje bi napadač opravdano mogao koristiti za povezivanje objavljene tablice s drugim pomoćnim izvorom; svaka bi grupa trebala uključiti najmanje *k* identičnih kombinacija generaliziranih atributa (tj. trebala bi predstavljati klasu ekvivalencije atributa). Tada bi skupove podataka trebalo objaviti samo nakon njihove podjele u takve homogene skupine. Atributi odabrani za generalizaciju u literaturi poznati su kao kvazi-identifikatori, jer bi njihovo poznavanje u čistom obliku omogućilo trenutačno utvrđivanje identiteta osoba čiji se podaci obrađuju.

[30] L. Sweeney. Ispreplitanje tehnologije i politike za očuvanje povjerljivosti. *Journal of Law, Medicine &Ethics*, 25, br. 2&3 (1997.): str. 98. – 110.

Mnogi eksperimenti utvrđivanja identiteta pokazali su slabost loše organiziranih tablica k-anonimnosti. To se na primjer može dogoditi zbog toga što su ostali atributi u klasi ekvivalencije identični (kao što je slučaj s klasom ekvivalencije španjolskih osoba čiji se podaci obrađuju u primjeru tablice A2) ili je njihova distribucija vrlo neuravnotežena s velikom učestalosti određenog atributa, ili zato što je broj zapisa u klasi ekvivalencije vrlo mali, čime se u oba slučaja pojačava vjerojatnost izvođenja zaključaka, ili zato što ne postoji značajna „semantička" razlika između čistih atributa klase ekvivalencije (npr. kvantitativna mjera takvih atributa mogla bi zapravo biti različita ali numerički vrlo bliska, ili bi mogli pripadati rasponu semantički sličnih atributa, npr. istoj razini kreditnog rizika ili istoj obitelji patologija), tako da skup podataka i dalje može propuštati veliku količinu podataka o osobama čiji se podaci obrađuju za napade povezivanja[31]. Ovdje je važno reći da kadgod su podaci raštrkani (na primjer određeno se svojstvo nekoliko puta pojavljuje na geografskom području), a prvim se sažimanjem ne mogu grupirati podaci s dovoljnim brojem pojavljivanja različitih svojstava (na primjer i dalje mali broj pojavljivanja nekoliko svojstava može se smjestiti u geografsko područje), potrebno je daljnje sažimanje atributa kako bi se postigla ciljana anonimizacija.

l-raznolikost

Polazeći od ovih zapažanja, tijekom godina predlagane su varijante k-anonimnosti i razvijeni su određeni inženjerski kriteriji za poboljšanje prakse anonimizacije pomoću generalizacije s ciljem smanjenja rizika od napada povezivanja. Oni se temelje na probabilističkim svojstvima skupova podataka. Posebno, dodano je daljnje ograničenje, naime da se svaki atribut u klasi ekvivalencije pojavljuje najmanje *l* puta, tako da napadač uvijek ima značajnu nesigurnost o atributima čak i uz prethodno znanje o određenoj osobi čiji se podaci obrađuju. To je isto kao da se kaže da bi skup podataka (ili potpodjela) trebao posjedovati minimalni broj pojavljivanja odabranog svojstva: ovo lukavstvo bi moglo ublažiti rizik od ponovnog utvrđivanja identiteta. To je cilj prakse anonimiziranja l-raznolikosti. Primjer ove prakse nalazi se u tablicama A4 (originalni podaci) i A5 (rezultat obrade). Kao što se vidi, pravilnim planiranjem/organiziranjem identifikatora lokacije i dobi pojedinaca u tablici A4, generalizacija atributa rezultira značajnim povećanjem nesigurnosti aktualnih atributa svake osobe čiji se podaci obrađuju u anketi. Na primjer, čak i ako napadač zna da osoba čiji se podaci obrađuju pripada prvoj klasi ekvivalencije, ne može dalje utvrditi posjeduje li osoba svojstvo X, Y ili Z, jer u toj klasi postoji najmanje jedan zapis (i u svakoj drugoj klasi ekvivalencije) s istim svojstvima.

[31] Treba naglasiti da se korelacije također mogu uspostaviti kada su zapisi podataka grupirani po atributima. Kada nadzornik podataka zna vrste korelacija koje želi provjeriti, može odabrati atribute koji su najmjerodavniji. Na primjer, rezultati ankete PEW nisu podložni napadima finog zrnatog izvođenja zaključaka i dalje su vrlo korisni za traženje korelacija između demografskih podataka i interesa (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx)

Serijski broj	Identifikator lokacije	Dob	Svojstvo
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

Tablica A4. Tablica s pojedincima grupiranim po lokaciji, dobi i trima svojstvima X, Y i Z

Serijski broj	Identifikator lokacije	Dob	Svojstvo
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

Tablica A5. Primjer 1-raznolike verzije tablice A4

t-bliskost:

Posebnim slučajem atributa unutar potpodjele koji su nejednako distribuirani ili pripadaju malom rasponu vrijednosti ili semantičkih značenja bavi se pristup poznat kao t-bliskost. Ovo je daljnje poboljšanje anonimizacije pomoću generalizacije i sastoji se od prakse raspoređivanja podataka kako bi se postigle klase ekvivalencije koje zrcale početnu distribuciju atributa u originalnom skupu podataka koliko god je to moguće. U ovu svrhu se koristi postupak od dva koraka na sljedeći način. Tablica A6 originalna je baza podataka koja uključuje čiste zapise o osobama čiji se podaci obrađuju, grupirane po lokaciji, dobi, plaći i dvije obitelji semantički sličnih svojstava, (X1,X2, X3) odnosno (Y1, Y2, Y3) (npr. slične klase kreditnog rizika, slične bolesti). Prvo se tablica *l-diversificira* s $l = 1$ (tablica A7), grupiranjem zapisa u semantički slične klase ekvivalencije i loše ciljane anonimizacije; zatim se obrađuje kako bi se u svakoj potpodjeli dobila t-bliskost (tablica A8) i veća varijabilnost. Zapravo, u drugom koraku svaka klasa ekvivalencije uključuje zapise iz obje obitelji svojstava. Valja primijetiti da identifikator lokacije i dob imaju različite zrnatosti u različitim koracima procesa: to znači da bi svaki atribut mogao zahtijevati različit kriterij generalizacije

kako bi se dobila ciljana anonimizacija, a to naprotiv od nadzornika podataka zahtijeva posebno planiranje i odgovarajući računalni teret.

Serijski broj	Identifikator lokacije	Dob	Plaća	Svojstvo
1	1127	29	30 000	X1
2	1112	22	32 000	X2
3	1128	27	35 000	X3
4	1215	43	50 000	X2
5	1219	52	120 000	Y1
6	1216	47	60 000	Y2
7	1115	30	55 000	Y2
8	1123	36	100 000	Y3
9	1117	32	110 000	X3

Tablica A6. Tablica s pojedincima grupiranima po lokaciji, dobi, plaćama i dvije obitelji svojstava

Serijski broj	Identifikator lokacije	Dob	Plaća	Svojstvo
1	11**	2*	30 000	X1
2	11**	2*	32 000	X2
3	11**	2*	35 000	X3
4	121*	> 40	50 000	X2
5	121*	> 40	120 000	Y1
6	121*	> 40	60 000	Y2
7	11**	3*	55 000	Y2
8	11**	3*	100 000	Y3
9	11**	3*	110 000	X3

Tablica A7. *l*-raznolika verzija tablice A6

Serijski broj	Identifikator lokacije	Dob	Plaća	Svojstvo
1	112*	< 40	30 000	X1
3	112*	< 40	35 000	X3
8	112*	< 40	100 000	Y3
4	121*	> 40	50 000	X2
5	121*	> 40	120 000	Y1
6	121*	> 40	60 000	Y2
2	111*	< 40	32 000	X2
7	111*	< 40	55 000	Y2
9	111*	< 40	110 000	X3

Tablica A8. *t*-bliska verzija tablice A6

Treba jasno naglasiti da se cilj generalizacije atributa osoba čiji se podaci obrađuju na tako učen način ponekad može ostvariti samo za mali broj zapisa, a ne za sve. Dobrom bi se praksom trebalo osigurati da svaka klasa ekvivalencije sadržava više pojedinaca i da ne ostaje mogućnost napada putem izvođenja zaključaka. U svakom slučaju, ovaj pristup od nadzornika podataka zahtijeva dubinsku procjenu dostupnih podataka uz kombinatornu ocjenu različitih opcija (na primjer, različite amplitude raspona, različita zrnatost lokacije ili dobi itd.). Drugim

riječima, anonimizacija pomoću generalizacije ne može biti rezultat grubog prvog pokušaja nadzornika podataka da zamijeni analitičke vrijednosti atributa u zapisu rasponima, jer su potrebni specifičniji kvantitativni pristupi – poput ocjenjivanja entropije atributa unutar svake potpodjele, ili mjerenja udaljenosti između distribucije originalnih atributa i distribucije u svakoj klasi ekvivalencije.