MNB-BME Együttműködés
2020/2021
Digitalizáció, mesterséges
intelligencia és adatkorszak Műhely

Csarnó Tamás Péter, Gulyás Gábor György

BANKI TRANZAKCIÓK ANONIMITÁSI KÉRDÉSEI

BUDAPEST, 2021

Tartalomjegyzék

Vezetői összefoglaló

A banki tranzakciók megosztása lassan általános jelenséggé válik, s egyre több startup kínál ezek az adatokra épülő szolgáltatásokat. Ezeket az adatokat csak úgy lehet megosztani és feldolgozni, ha az megfelel a hatályos szabályozási környezetnek, többek között az európai Általános Adatvédelmi Rendeletnek (GDPR). A GDPR szigorú feltételeket támaszt a személyes adatok védelmét illetően, amelyek alól kivételt képez, ha az adatok anonimizáltak. Ekkor ugyanis már nem állítható helyre az adatalanyok személyazonossága, s nem számítanak személyes adatoknak. Ezen adatokat szabadon fel lehet használni jelentősebb adatvédelmi korlátozások nélkül.

Azonban a tranzakciós adatok megfelelő anonimizálása nem egyértelmű feladat. Egyrészt hiába nagyméretűek ezek az adatbázisok, az egyes felhasználói rekordok ennek ellenére egyediek és beazonosíthatók. Másodsorban, az adatstruktúra sajátosságai alapvetően felülírják a klasszikus anonimizálás lehetőségét: a tranzakciók számossága ellehetetleníti az adatok törlésen és általánosításon alapuló anonimizálását. Harmadrészt, a kapcsolatrendszer jellegű felépítés (például ki-kinek utalt) a táblázatos adatokhoz képest plusz lehetőséget biztosít az újra-azonosításra.

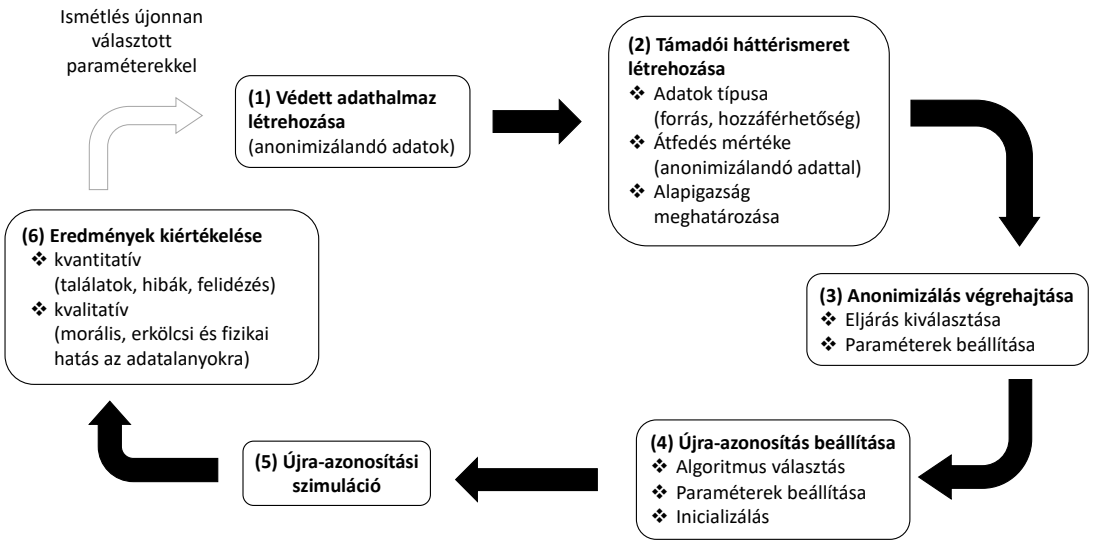
Jelen tanulmány célja mélységében feltárni ezeket a kihívásokat, és egy új, kifejezetten a tranzakció anonimizálás problémájához illő módszertannal támpontokat adni a megfelelő anonimizálási módszer kiválasztásához. Ennek érdekében a tanulmány újszerű megközelítéssel vizsgálja a tranzakciós adatok anonimitását: nem hagyományos, például relációs-táblázatos adatokként, hanem gráf struktúrában kezeli azokat (vagy másképp fogalmazva hálózatokként).

Az anonimizálási eljárások hatékonyságát mindig a legkorszerűbb újra-azonosítási eljárásokhoz kell viszonyítani. Ez nem csupán logikus megközelítés, hanem így felelhetünk meg a GDPR elvárásának is: az anonimitási eljárás megválasztásakor „*összes objektív tényezőt figyelembe kell venni, így például az azonosítás költségeit és időigényét, számításba véve az adatkezeléskor rendelkezésre álló technológiákat, és a technológia fejlődését*” [1] (GDPR, 26. cikk).

Az újra-azonosítási eljárások jellemzően abból indulnak ki, hogy valamilyen háttérinformáció (háttérismeret) segítségével az anonimizált adatban bizonyos esetekben

a rekordok identitása helyreállítható. A háttérismeret lehet akár néhány személyre vonatkozó kiegészítő információ, de lehet akár egy komplett adatbázis is. Ez utóbbi jellegében lehet hasonló a tranzakciós adatokhoz (például kapcsolati háló), de ez nem mindig szükségszerű. Jelen esetben olyan újra-azonosítási algoritmusokat vizsgálunk, amelyek valamilyen alternatív forrásból származó kapcsolati hálózatot használnak fel az újra-azonosítás végrehajtásához (például közösségi hálózatok, levelezés és hívás kapcsolatok).

Annak érdekében, hogy széleskörűen felmérjük az anonimitás megsérülésének kockázatát, iteratív megközelítésű módszertant alkalmazunk, amelyben a vizsgált anonimizálási eljárást többféle támadással veti össze. A végeredményként előálló eredményhalmaz elemei (például a támadási kísérletek sikerességi és hiba arányai) döntéselőkészítésre, következtetések megalkotására (például adatvédelmi hatásvizsgálatokhoz) hasznosak lehetnek. A módszertan lépéseit mutatjuk be az I. ábrán.



Ábra I. Módszertan tranzakciós és hálózatos (gráf struktúrájú) adatok anonimizálási eljárásainak vizsgálatára.

Kulcsfontosságú, hogy az anonimizálási eljárás ellenálló képességét különféle képességi szintű (erősségű) rosszindulatú féllel szemben legyen vizsgálva. Erre a legtöbb esetben nincs lehetőség, ugyanis bár az anonimizálandó adat rendelkezésre áll, de a támadó képességeinek megfelelő háttérismeret nem. Áthidalhatjuk ezt a szintetikus adat

generálási eljárásokkal, melyek egyetlen adathalmazból képesek előállítani egy kellően hihető háttérismeret és anonimizálandó adathalmaz párt.

A vizsgálatainkban mintavételezéssel állítjuk elő az adatok az anonimizálandó adatból, paraméterrel meghatározva a csúcs és él átfedés mértékét. Ezzel tetszőlegesen beállítható, hogy milyen erősségű támadót szeretnénk vizsgálni. Mivel az adatokat szintetikus módon hozzuk létre, ismerjük azokat a csúcsokat, melyek mind az anonimizálandó adatban és a támadó háttérismeretében benne vannak, az ezek közötti megfeleltetéseket is ismerjük. Ez az ún. alapigazság, amellyel újra-azonosítási szimulációk futtatása után már ki fogjuk tudni értékelni az eredményeket.

Az anonimizációs eljárás alkalmazása után le tudjuk futtatni az újra-azonosítási eljárást ezzel szemben. Az eljárás végeredménye egy hozzárendelési halmaz, amelyben az algoritmus megadja, hogy mely azonosított háttérismereti csúcsoknak felelnek meg az anonimizált adathalmazban lévő csúcsok. Ezek alapján meghatározható, hogy az alapigazság szerint rendelkezésre álló csúcsok közül mekkora arány volt képes megtalálni (felidézés), illetve a találatok hány százaléka volt helyes.

A módszertan működését és hatékonyságát az Enron e-mail-ezési adathalmazon mutatjuk be egy komplex esettanulmány formájában, amely 36 ezer személy levelezése alapján készült. A körülbelül félmillió levelet érintő adathalmazt az amerikai Federal Energy Regulatory Commission tette közzé, amikor vizsgálatot folytatott le az Enron céggel szemben 2001-ben [30].

A vizsgálat során három különböző képességű támadót vizsgálunk, amelyek eltérő erősségű háttérismerettel rendelkeznek. A gyenge támadó háttérismeretében mindössze 4 306 csúcs van, amely az anonimizált adathalmazban is szerepel, míg a közepesen erős támadóéban 12 119, az erős támadóéban pedig 21 507 ilyen csúcs van.

Ha a vizsgálatot egy adatvédelmi hatásvizsgálathoz folytatjuk le, vélhetően egy bizonyos anonimizálási eljárást fogunk egy javasolt beállítással vizsgálni, de a teljesség kedvéért az esettanulmányban több anonimizálási eljárást is összehasonlítunk. A legegyszerűbb, *Switch(k)* nevű eljárás csak véletlenszerűen áthelyezi az élek egy részét. A további eljárások egy-egy klasszikus anonimizálási eljárás mintájára alakítják át az adathalmazt. A *k-DA* eljárást a fokszámeloszlást úgy alakítja át, hogy minden fokszám értékhez legalább k darab csúcs tartozzon (ez a k-anonimitás mintájára működik), míg a

DP(ε) eljárás a differenciális adatvédelem elvének megfelelően alakítja át a gráfot. Az eljárások paraméterei úgy választottuk meg, hogy a szakirodalmi ajánlások alapján erős anonimizálást kapjunk [22]. Ezen a beállítások mellett a gráfok struktúrája még használható marad, így ebből a szempontból az anonimizálás megfelelő.

	Nar algoritmus [2]	**Blb algoritmus [10]**
Switch(k)* *(k=10)	25,88%	35,07%
k-DA* *(k=50)	30,65%	39,04%
DP(ε) **($\varepsilon = 50$)**	22,28%	26,67%

Táblázat I. Erős Támadó újra-azonosítási eredményei (felidézés): mivel a csúcsok jelentős hányadás sikerült újból azonosítottá tenni, további anonimizálási beállításokat kell keresni, amelyeknél az adatok hasznossága kevésbé sérül, de az erős támadóval szemben is képesek védelmet nyújtani.

A gyenge támadóval szemben ezek az anonimizálási beállítások védelmet nyújtanak, mivel a támadó nem volt képes jelentős, néhány százaléknál nagyobb felidézést elérni. Azonban közepes és erős támadó esetén már sikerrel járt, melyre példaként az erős támadó felidézési eredményeit mutatjuk be az I. táblázatban. Ezért egy újabb anonimizálási beállítást kell keresni, vagy másik anonimizálási eljárást kell kipróbálni.

Mivel a leghatékonyabb védelmet a *DP(ε)* nyújtja, ennek a paraméterét állítottuk erősebb védelmi szintűre ($\varepsilon = 25$), amely már megfelelőnek bizonyult: az adatok hasznossága továbbra is elfogadható szinten maradt, de semelyik újra-azonosítási algoritmus nem tudott 2-3%-nál nagyobb felidézést elérni.

A fentebb bemutatott esettanulmány hatékonyan bemutatja, hogy valós adatokon hogyan alkalmazhatjuk a tanulmányban elsőként bemutatott módszertant a megfelelő adatvédelmet nyújtó anonimizálási eljárás kiválasztásában.

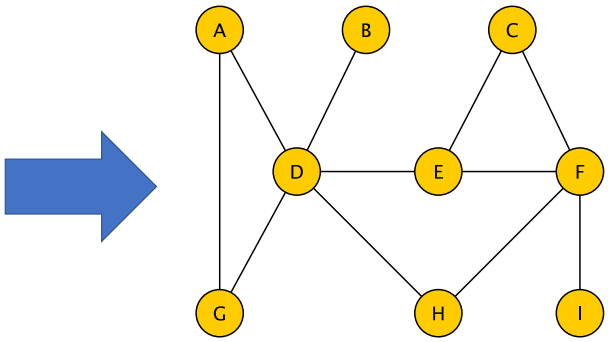
1 Bevezetés

Ahogy a legtöbb szolgáltatás esetén szükséges vagy hasznos lehet adatokat megosztani, ez alól a banki adatok sem kivételek. Bár a felügyeleti szervek irányába történő adatok megosztása külön esetet képviselnek, általánosságban elmondható, hogy a ma felfutásban lévő pénzügyi digitalizációs törekvések egyik jelentős eleme a tranzakciós adatok monetizálása lesz.

A tranzakciós adatok jellemzően magánszemélyek közti pénzmozgásra, magánszemélyek és vállalatok, illetve vállalatok közti pénzügyi eseményekre vonatkoznak. Ezeket az adatokat az informatikai rendszerek a relációs tárolási módból fakadóan jellemzően táblázatos formában tárolják. Ebből fakadóan felmerülhet, hogy ezeket az adatokat ugyanilyen formában kínálja fel egy pénzügyi szolgáltató megosztásra, miután az azonosító adatokat eltávolították, esetleg megkísérelték az adatok anonimizálását.

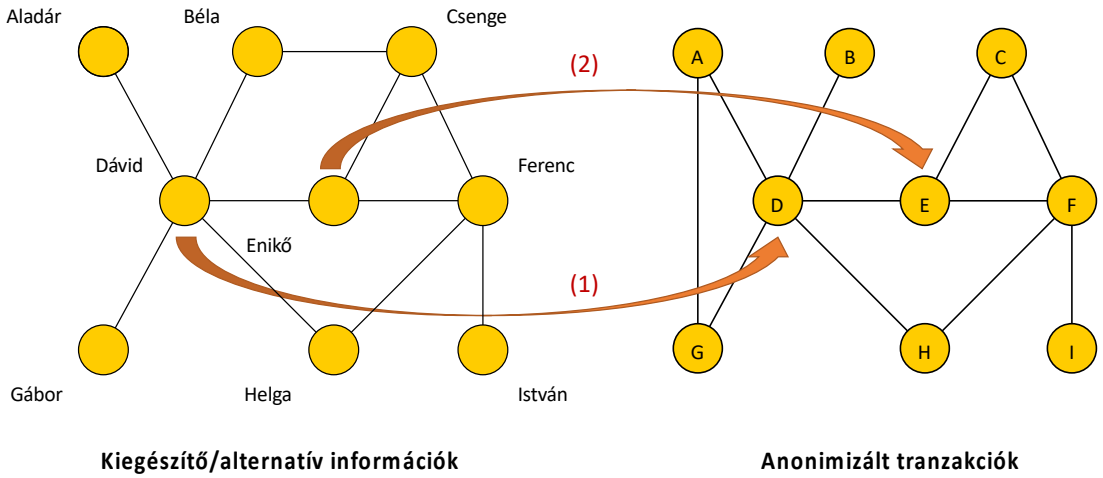
Jelen tanulmányban megmutatjuk, hogy a tranzakciós adatok anonimizálása nehéz feladat, ugyanis a tranzakciót végző entitások közötti összefüggéseket kihívást jelenthez eredményesen semlegesíteni, úgy, hogy közben az adatbázis felhasználhatósága is megmaradjon. Emögött elsősorban az áll, hogy egy tranzakciós adatbázis egy nagyméretű hálózattá (gráffá) konvertálható, és hiába távolítjuk el az azonosítókat, az entitásokat körülvevő strukturális információ önmagában lehetővé teszi az egyes entitások beazonosítását [2]. (A továbbiakban a gráf és hálózat fogalmakat kölcsönösen felcserélhetően használjuk.) Azt is megmutatjuk, hogy mintavételezéssel sem lehet megfelelő szintű adatvédelmet biztosítani a tranzakciókat végrehajtó személyek számára [9].

	A	B	C	D	E	F	G	H	I
A				X			X		
B				X					
C					X	X			
D	X	X			X		X	X	
E			X	X		X			
F			X		X			X	X
G	X			X					
H				X		X			
I						X			



Ábra 1. A tranzakciók táblázata (relációs táblája) értelmezhető egy mátrix szomszédossági mátrixa ként is (balra). A szomszédossági mátrixból előálló gráf struktúráját klasszikus gráf- és hálózatelméleti eszközökkel is feldolgozhatjuk.

A struktúra felhasználása a következőképpen történhet. Amikor a pénzintézet publikálja a tranzakciós adatokat, megfelelő anonimizáció esetén rosszindulatú felek nem tudnak közvetlenül visszaélni az adatokkal, mert nem tudják az egyes tranzakciókat természetes vagy jogi személyekhez kapcsolni. Azonban a struktúra kihasználásához meg lehet próbálni találni egy olyan kiegészítő adatbázist (ún. háttérinformáció), amiben vannak már nevek vagy azonosítók, hogy majd a két adatbázis összevetése során ezeknek segítségével a névtelen rekordokat azonosíthassuk.



Ábra 2. Egy rosszindulatú harmadik fél próbálkozása a jobb oldalon látható hálózat identitásainak megismerésére. Célja a névtelenül publikált hálózattal együtt megosztott érzékeny információk korlátlan felhasználása az anonimizálás előtti identitások helyreállítása által. Ehhez először globálisan kiugró struktúrájú pontokat keres (1), mint például Dávid. Az ő fokszáma (kapcsolatainak száma) az egész hálózatban egyedi, ami vélhetően teljesül az anonimizált adatban

is. A támadó így Dávid kapcsolatba hozza D-vel. Majd ezek környezetében lokálisan kiugrókat (2), mint például Enikő, akinek a fokszáma nem egyedi globálisan, de Dávid környezetében igen.

Az 1. ábrán jobb oldalon látható egy tranzakció adatbázisból létrehozott anonim hálózat. Ebben a hálózatban nem ismertek az egyes csomópontok eredeti adatalanyát azonosító információi, de más, érzékeny információk ismertek lehetnek. Egy rosszindulatú harmadik fél, ha hozzájut ehhez az információhoz, megpróbálkozhat annak de-anonimizálásával, vagy más néven újra-azonosításával. (Ezen két fogalmat is csereszabatosan használjuk a későbbiekben.)

Azt a folyamatot hívjuk így, amikor egy anonimizált (vagy álnevesített, ld. 2. fejezet) adathalmaz entitásaihoz megpróbáljuk helyreállítani az eredeti azonosítókat. Amennyiben az anonimizálás során helyesen jártunk el, ez csak további információforrások bevonásával lehetséges. A támadó fél például próbálkozhat hasonlóan strukturált, hálózat alapú adatokforrások bevonására. Ma hasonló jellegű adathalmazokból számos különféle típus állhat rendelkezésre, ami alkalmas lehet a tranzakciós adatok de-anonimizálására. A klasszikus közösségi hálózatokon kívül hasonló gráf struktúrát lehet létrehozni hívás, email és egyéb kommunikációs adatbázisokból, de a szakirodalomban olyan algoritmusok is ismertek, amelyek az egy helyen eltöltött idő alapján képesek ezt megtenni [11].

Tegyük fel, hogy a rosszindulatú fél talál mondjuk egy olyan közösségi hálózatot, ami feltételezése szerint nagyban átfed a tranzakciós adatbázissal. A példa kedvéért az átfedés teljes, de a későbbiekben látni fogjuk, hogy akár egész kis átfedés esetén is sikeres lehet egy ilyen de-anonimizációs támadás.

A támadás elején nem tudjuk melyek azok a gráf csomópontok, amelyek mindkét hálózatban szerepelnek, nemhogy az ezek közötti megfeleltetéseket. Ezért a de-anonimizálási eljárás első fázisában (az ún. seed, vagy magyarul a magvetési fázis) olyan csomópont párokat kell keresnünk, amelyek a hálózat egészében kiugró és egyedi strukturális tulajdonságokkal rendelkeznek. Az 1. ábrán lévő példában ilyenek a Dávid/D csomópontok, amelyek mindkét hálózatban egyedien azonosíthatóak fokszámuk (szomszédjaik számossága) alapján. Ezeknél élhetünk azzal a feltételezéssel, hogy a két csomópont ugyanazt a személyt jelöli, és a D-hez tartozó pénzügyi adatok a Dávid nevű személynek az adatai. Azonban példánkban ilyen módon további entitások nem

azonosíthatók be. Megjegyezzük, hogy bizonyos támadó algoritmusok kihagyják ezt a fázist, azaz nem igényelnek kezdeti párosításokat.

Megfelelő számú kezdeti találat esetén indul el a második fázis (az ún. propagation, vagy terjedési fázis), amikor a kezdeti találatok mentén bővítjük tovább a párosításokat. Például Enikő fokszáma három, ami nem egyedi a gráf egészében. Ezért önmagában nem lehet beazonosítani; az anonimizált hálózatban lévő E akár Csengének vagy Enikőnek is lehetne a megfelelő párja. Azonban tudjuk, hogy Enikő közvetlen szomszédja Dávidnak, ami alapján tudjuk, hogy az lesz az Enikő anonimizált hálózatbeli megfelelője, ami szintén három fokú, és szomszédja D-nek (Dávid megfelelőjének). Ez alapján megfeleltetést tudunk létrehozni Enikő és E között. Majd ezt az eljárást kell folytatnunk, amíg találunk újabb és újabb megfeleltetési lehetőségeket.

Bár a valóságban a rendelkezésre álló információforrások kevésbé vannak átfedésben az anonimizált adatokkal, és a struktúra is jelentősebben eltérhet, azonban a későbbiekben látni fogjuk, hogy az algoritmusok nem érzékenyek a zajra, és egész kicsi átfedéssel is képesek elboldogulni.

Jelen tanulmány célja módszertant adni az olvasó kezébe, hogy megismerve a hálózati struktúrára épülő de-anonimizációs algoritmusokat, az olvasó fel tudja ezeket használni a vonatkozó anonimizálási eljárások minősítésére. Az Általános Európai Adatvédelmi Rendelet (GDPR) [1] az Európai Unió és az Európai Gazdasági Térség valamennyi országában kötelezően érvényben lévő adatvédelmi szabályozás, amely megköveteli a természetes személyekre vonatkozó információk (személyes adatok) szigorú védelmét.

A GDPR többféle technikai adatvédelmi megoldást támogat, amely közül az anonimizálás az egyik legerőteljsebb eljárás. Ugyanis megfelelően végrehajtott anonimizálás által az adat elveszíti személyes adat jellegét, és már nem vonatkozik rá a továbbiakban a GDPR (26. cikk, [1]). Azonban az anonimizálást végrehajtó adatkezelőnek feladata megvizsgálni, hogy valamennyi, ésszerű kereteken belül feltételezhető rosszindulatú fél esetén a de-anonimizációs támadás sikerülhet-e. Akkor tekinthető az anonimizálási eljárás elfogadhatónak, ha a rosszindulatú fél nagy valószínűséggel kudarcra van ítélve (szintén 26. cikk, GDPR [1]).

Tanulmányunkban ismertetett módszertan lehetőséget ad különböző erősségű rosszindulatú felek (támadók) vizsgálatára, amely segítségével vizsgálhatóak a hálózatok

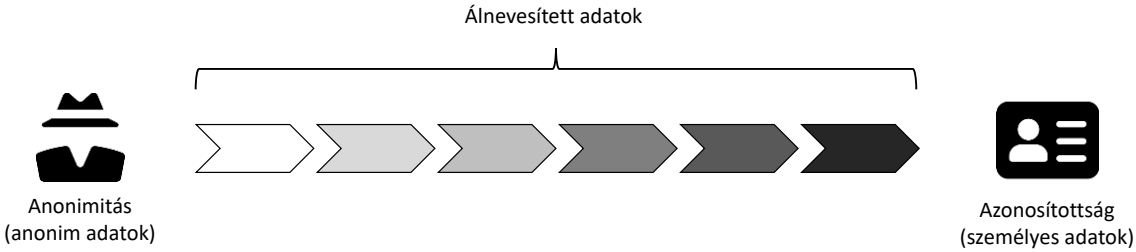
anonimizálási és de-anonimizálási eljárásai. Ezzel a módszertannal meg lehet keresni azt a köztes állapotot az adatok használhatósága és adatvédelmi szintje között, amikor az anonimizálás még nem lehetetleníti el az adatok felhasználását, de a támadó sikerességének valószínűsége ésszerű keretek közé szorítható.

2 Az anonimizálásról

Az anonimizálás célja bizonyos adatok, rekordok módosítása olyan módon, hogy megszűnjön a kapcsolat helyre nem állítható módon azok és az eredeti adatalany között. Adatbázisok, adatbázisokból kiexportált részek anonimizálására szükség lehet a későbbi megőrzés céljából, vagy azért, mert szándékunkban áll megosztani azt, és ezt anonimizálás nélkül túlságosan kockázatos lenne megtenni.

2.1 Személyes adatok és anonimitás

Az online közösségi hálózatok és a nagy adat jellegű adatbázisok (angolul az ún. big data) megjelenésével rengeteg személyes adat kering az interneten. A felhasználók személyes adatai rendkívül értékesek. A felhasználók internetes tevékenységeikről és szokásaikról, preferenciáikról szóló információkra nagy kereslet van többek között célzott hirdetésekhez, alkalmazásfejlesztéshez, adatbányászathoz. Ez a tendencia aggályt keltett a felhasználókban azzal kapcsolatban, hogy hogyan gyűjtik, tárolják és használják fel a személyes adataikat [9].



Ábra 3. Adatok három azonosítási szintje. Jellemzően egyszerűbb több információt elárulni egy profilról, mint visszavonni ezeket; erre utalnak a jobbra mutató nyilak, amelyek különböző szintű álnevesített adatokat jelölnek, melyek azonosítása eltérő komplexitású feladat.

A GDPR definíciója alapján a **személyes adatnak** tekintendő minden olyan információ, amely valamely azonosított vagy azonosítható élő személlyel kapcsolatos [1]. Mindazon információk, amelyek összegyűjtése egy bizonyos személy azonosításához vezethet, ugyancsak személyes adatnak minősülnek. Személyes adatnak számít többek között a természetes személy neve, személyazonosító száma, illetve valamennyi testi, fiziológiai, genetikai, szellemi, gazdasági, kulturális vagy szociális azonosságára vonatkozó információ is.

Az adatok (egy adatbázis rekord oszlopa, attribútuma) az azonosíthatóság szempontjából a következő három kategóriába sorolhatók:

1. **Közvetlen azonosítók**: ezek az adatok egyértelműen azonosíthatják az egyéneket. Ide tartozik az illető neve, személyi azonosító száma, lakcíme.
2. **Kvázi-azonosítók**: azok az attribútumok, amelyek értékei együttesen azonosíthatják az egyént. Kvázi azonosító lehet például egy illető, születési dátuma, életkora vagy neme.
3. **Érzékeny adatok**: például ismert betegség, fizetés. Ezeket az adatokat kívánjuk hasznosságuk miatt megosztani, de mindamellett meg is óvni az illetéktelen felhasználással szemben.
4. **Egyéb adatok**, amelyek nem tartoznak a fenti kategóriákba.

Egy adat önállóan tekintve akkor számít ebből az aspektusból személyes adatnak, ha közvetlenül beazonosítható általa az érintett természetes személy (közvetlen azonosítók), vagy ha közvetetten (kvázi-azonosítók). Azonban az adatokat nem csak önállóan vizsgáljuk, hiszen egy rekordban más adatok is lehetnek, amelyek önállóan nem azonosíthatók, és nem is hasznosíthatók. Például a pénzügyi adatok akkor monetizálhatók jól, ha további adatokat (életkor, lakhely, foglalkozás) is kapcsolunk hozzájuk, amelyek azonban kvázi-azonosítóként is funkcionálhatnak. Ezért az ilyen adatoknak (3-4. kategória) az azonosíthatóságát a rekordban szereplő közvetlen és kvázi-azonosítók alapján kell vizsgálni.

A személyes adatok védelme anonimizálási eljárásokkal elérhető, ekkor úgy módosítunk az adaton, hogy az adatvédelmi kockázatokat minél inkább minimalizáljuk. **Anonim adatnak** minősül az olyan adat, amelynek személyes jellege már nem állítható helyre, az adat átalakítása nem visszafordítható.

A GDPR megkülönbözteti még az **álnevesített adatokat**, amelyek önmagukban nem képesek egy személy közvetlen azonosítására, de további információ felhasználásával már közvetetten kapcsolatba lehet hozni egy személlyel. Az álnevesített adatoknak sok szintje lehet attól függően, hogy mennyire van elválasztva a személyes adattól (lásd: 3. ábra). Az anonimizálással ellentétben az álnevesítés visszafordítható az adatban lévő kvázi-azonosítókhoz a megfelelő külső adatforrás bevonásával.

Míg személyes adatokból egy rosszindulatú támadó közvetlenül ki tudja nyerni az érzékeny információkat, álnevesített adatok esetén már szükséges valamennyi szaktudás és háttéradat a sikeres újra-azonosításhoz. Anonim adatok esetén az újra-azonosítás komplexitása a legmagasabb.

2.2 **Anonimizálási eljárások**

Jelenleg különböző anonimizálási gyakorlatok és technikák léteznek. Kitérünk azokra a főbb szempontokra, amelyeket figyelembe kell venni egy adott technika alkalmazásakor. A 29-es munkacsoport anonimizálási eljárásokról szóló véleménye [15] alapján az alábbi három fő szempontot célszerű mérlegelni:

a) **Kiválasztás**ról (singling out) akkor beszélhetünk, ha egy rosszindulatú fél sikeresen be tud azonosítani egy adott személyhez tartozó rekordot az adathalmazon belül.

b) **Összekapcsolhatóság** (linkability) legalább két rekord összekapcsolásának képessége, amely ugyanahhoz az érintetthez vagy érintettek csoportjához tartozik. Nem szükséges pontosan beazonosítani az adott személyeket.

c) **Következtetés** (inference) ami annak a lehetőségét jelenti, hogy nagy valószínűséggel jelentősen új információ következtethető ki az attribútumok értékeiből.

Ezen három fő szempontnak megfelelő anonimizálási technika elegendő védelmet nyújthat a rosszindulatú fél által legvalószínűbben alkalmazott újra-azonosítási eljárásokkal szemben. Noha meg kell jegyezzük, hogy a GDPR anonimitással kapcsolatos elvárásai között a kiválasztás és összekapcsolhatóság kritériumnak való megfelelésen van a fő hangsúly. Mint azt látni fogjuk, önmagában egyik anonimizálási technika sem mentes a hiányosságoktól, ezért célszerű több technikát együttesen alkalmazni, így erősebb adatvédelmi garanciákat biztosíthatunk. Általánosságban véve az anonimizálási technikák módszer szerint két különböző csoportba sorolhatók. Az első a véletlenítésen, a másik az általánosítás elvén alapul.

2.2.1 Véletlenítés

A véletlenítés adattorzítási technikákat foglal magába, amely után az adatok kevésbé egyeznek meg egy az egyben az eredetivel. A technika célja az adat és az adatalany szoros

egyezéseinek a feloldása, így azok nehezebben köthetőek egy adott személyhez. A véletlenítés tehát védelmet nyújthat a következtetési kockázatok ellen.

A **zajhozzáadás** technika ebbe a csoportba sorolható. A zajhozzáadás célja az adatok módosítása véletlen értékekkel oly módon, hogy azok kevésbé feleljenek meg a valóságnak, mindemellett megtartva az általános eloszlást. A technika megfelelő alkalmazásakor a rosszindulatú harmadik fél kisebb valószínűséggel lesz képes egy egyén azonosítására, illetve nehezen kimutatható, hogy az eredeti adat módosítva lett. A zajhozzáadás a következtetési támadások sikerességét rontja. Bár a zajhozzáadás megnehezíti a személyes adatok kinyerését, ezt a technikát gyakran egyéb anonimizálási eljárással érdemes kombinálni, például egyértelmű azonosítók és kvázi-azonosítók eltávolításával.

A **permutáció** technika a zajhozzáadáshoz hasonlóan működik, tekinthető annak egy speciális változatának is. Ez esetben egy táblázatos adatban található attribútumok értékeinek összekeveréséből áll. Ennek hatására az adatok tartománya és eloszlása nem változik, viszont az attribútumok közötti logikai kapcsolatok és statisztikai összefüggések megszűnnek. A zajhozzáadáshoz hasonlóan önmagában a permutáció sem feltétlenül biztosítja az anonimizálást, és mindig össze kell kapcsolni az egyértelmű azonosítók és kvázi-azonosítók eltávolításával.

A **differenciális adatvédelem** a véletlenítési technikák csoportjába tartozik, de a korábbiaktól eltérő módszert alkalmaz [19]. Ennek a módszernek a célja egy adott adathalmazhoz olyan mértékű zaj hozzáadása, ami szükséges az adatvédelmi célok eléréséhez, mindeközben az egyes személyre vonatkozó információknak hihetően letagadhatónak kell lenniük. A zaj hozzáadása során az adathalmazra jellemző statisztikák pontosak maradnak.

2.2.2 Általánosítás

Az anonimizálási technikák második csoportja az általánosítás. Ez a módszer azt jelenti, hogy az érintettek attribútumait egy bizonyos stratégia mentén általánosítják, elnyomják, törlik vagy felhígítják míg az anonimitási kritériumoknak meg nem felelnek. Az általánosítás védelmet nyújthat a kiválasztás kockázatával szemben, nem minden esetben teszi lehetővé a hatékony anonimizálást.

A **k-anonimitás** az általánosítás technikák közé sorolható. [6] A k-anonimítás célja megakadályozni az egyének kiválasztását úgy, hogy az egyéneket legalább k főből álló csoportokba rendezzük. Egy csoporton belüli kvázi-azonosító attribútumok értékei általánosítva vannak olyan mértékben, hogy minden egyén azonos értékkel rendelkezzen. A személyes adatokat eltérő módon lehet általánosítani, például születési dátumokat hónaponként vagy évenként lehet általánosítani míg más numerikus adatokat intervallumértékekkel általánosíthatók. A módszer feltétele még az egyértelmű és kvázi-azonosító adatok eltávolítása az adathalmazból. A k-anonimitás védelmet nyújthat a kiválasztással szemben, mivel ugyanazokat az attribútumokat most már k felhasználó osztja meg egymással, többé nincs lehetőség egy egyén kiválasztására. A k-anonimitási modell fő hiányossága, hogy nem akadályozza meg a következtetési támadások egyik típusát sem.

L-diverzitás és T-közelítés

Az érzékeny adatok kiszivárogtatása nagyobb probléma, mint maga az újra-azonosítás. K-anonimitás egyik hiányossága az, hogy következtetéssel továbbra is ki lehet nyerni érzékeny adatot akár de-anonimizálás nélkül is. A k-anonimitáshoz hasonlóan az l-diverzitás is könnyen kiszámítható miután kiválasztottuk az érzékenynek számító adatokat és a kvázi-azonosítókat az adathalmazunkban.

A további determinisztikus következtetési támadások megakadályozása érdekében az **l-diverzitás** [17] nemcsak azt írja elő, hogy minden egyén megkülönböztethetetlen legyen kellően sok más egyéntől, hanem azt is, hogy az egyes csoportokon belül legalább *l* különböző érzékeny attribútum szerepeljen. Míg az l-diverzitás erősebb védelmet nyújt, mint a k-anonimitás, az adat felhasználhatóságát nagyobb mértékben rontja.

Abban ez esetben, ha az attribútumértékek jól oszlanak el, az l-diverzitás hasznos módszer az adatok következtetési támadásokkal szembeni védelmére, viszont, ha az attribútumok egyenlőtlenül oszlanak el akkor továbbra is előfordulhat információszivárgás. Ez az alapgondolata a **t-közelség**nek [18]. Ez a technika szigorúbb megkötéseket támaszt, mint az l-diverzitás, mivel elvárja, hogy minden egyes értéknek az egyes attribútumok kezdeti eloszlásának megfelelő számban kell szerepelnie. Mivel ez a módszer további megkötéseket támaszt, tovább ronthat az adat felhasználhatóságán.

2.2.3 Anonimitás és felhasználhatóság

Mint azt láthattuk, személyes adatok anonimizálásával megszüntethetőek vagy legalább csökkenthetőek az adatvédelmi kockázatok. A fenti módszerek alkalmazása során sajnos elkerülhetetlen az információveszteség. Zaj hozzáadása esetén szándékosan rontunk az adatok pontosságán, személyes adatok általánosítása esetén bizonyos attribútumok törlődnek, vagy egyes demográfiai adatok módosulhatnak. Az adatok teljes megváltoztatása nagyon egyszerű és hatékony módszer lenne a vele kapcsolatos összes adatvédelmi kockázat megszűntetésére, de ez egyúttal az adathalmaz értékét is tönkre tenné.

Az anonimizálás fő kihívása tehát az adatok néhány hasznos tulajdonságának megőrzése a kockázatok korlátozása mellett. Általában mindkét tulajdonság csak a másik kárára növelhető, nem létezik olyan ideális megoldás, ahol maximális adatvédelem mellett teljes mértékben felhasználható maradna az adathalmaz. A két véglet közötti átmenetet az anonimizálási algoritmusok paramétereivel szabályozhatjuk, mint ahogy például a k-anonimitás esetén a k értékével megadhatjuk az ekvivalencia csoportok méretét, ahol a kvázi-azonosító jellemzők minden felhasználóra megegyeznek.

2.3 Anonimizálás nehézségei a nagy adatban

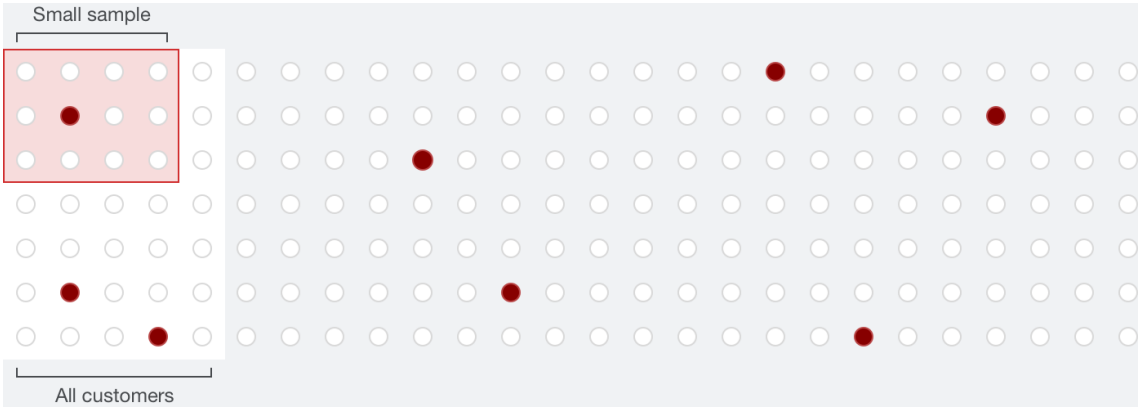
Az elmúlt évtizedben világszinten jelentősen megnőtt a személyes adatok gyűjtése és tárolása. Azonban a pénzügyi és orvosi szolgáltatásoktól összegyűjtött adatok felhasználhatóak számos visszaélésre is. Az adatok megosztása, vagy nyilvánossá tétele előtt azt anonimizálni szükséges az emberek személyes adatainak védelmének céljából.

Ennek ellenére mégis számos, elvileg anonim adathalmazt hoztak nyilvánosságra az utóbbi időben, amelyekről kiderült, hogy de-anonomizálhatóak. Sok esetben ezek az adathalmazok anonimizálása annyiból állt, hogy eltávolították a közvetlen azonosítókat (például nevet, e-mail címet, személyi azonosító számot), amely valójában csak álnevesítésnek felel meg a GDPR szempontjából. Az első prominens példa az volt, amikor az 1990-es évek végén Latanya Sweeney képes volt újraazonosítani William Weld kormányó orvosi adatait. Az újraazonosításhoz elegendő volt a kormányzó ZIP kódja, születési dátuma és neme [3]. Frissebb példa 2017-ből, amikor német újságírok nemrégiben sikeresen újraazonosították egy német képviselő böngészési előzményét egy anonimizált adathalmazból [12].

Alapvetően két (téves) elképzelés a meghatározó a nagy adatok publikálása során. Az egyik arra épít, hogyha csak az adatok részhalmazát publikáljak, akkor az hihetően letagadhatóvá teszi az egyes személyek jelenlétét a publikált adatban. A másik elképzelés szerint ezzel összefüggésben arra épít, hogy a nagyobb méretű adathalmazban nehezebben megtalálhatóak egyes személyek. Azonban a nagy adatra jellemző, hogy nem csak a rekordok száma nő meg, hanem az attribútumoké is. Ez azonban a sorok számának növekedése ellenére is könnyebben azonosíthatóvá teszi a benne lévő természetes személyeket, sőt, kifejezetten megnehezíti a hatékony anonimizálást.

2.3.1 Mintavételezésen alapuló „anonimizálás”

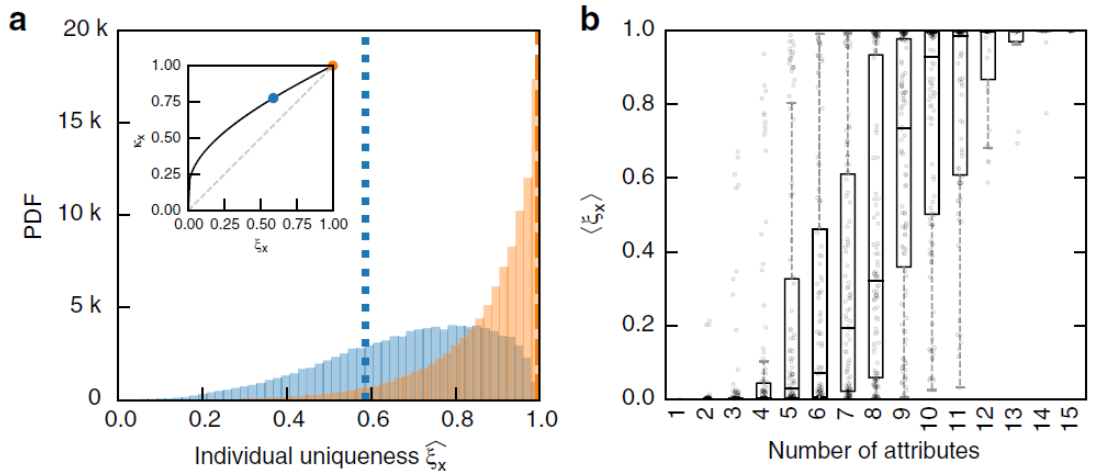
Védekezésképp a vállalatok gyakran az adathalmaz mintavételezett részét publikálják, ami annyit jelent, hogy a teljes adathalmaz megosztása helyett az adatok egy kisebb részét, néhány ezer ember adatait osztják meg. Az az elképzelésük, hogy a mintavételezéssel csökkenthetőek az adatvédelmi kockázatok, és az egyes személyekre vonatkozó információk felfedése újra-azonosítás által hihetően letagadható. Tehát, ha például egy egészségügyi adatokat tartalmazó adathalmaznak csak az 1%-át osztják meg, és a megosztott adathalmazban megtaláljuk a keresett személy demográfiai adatainak megfelelő rekordot, nem lehetünk biztosak abban, hogy az adatok valóban a keresett személyhez tartoznak. Könnyen lehet, hogy az adathalmaz 99%-ból valaki máshoz tartozik, aki a keresett személlyel megegyező adatokkal rendelkezik.



Ábra 4. minden pont egy egyénnek felel meg. Az egyének közül sokan azonos demográfiai adatokkal rendelkeznek (piros pont), míg másnál különböző. Az összes pont a teljes populációt jelképezi, aminek kis része az adathalmaz (pl. utalásokat tartalmazó adatbázis). A teljes adathalmaz csak kis részét publikálják (piros téglalappal jelölve) [21].

A 4. ábrán szemléltetjük a mintavételezés elvét. Ha valaki hozzáfér a mintavételezett adatokhoz (ami az eredeti adathalmaz egy kis részét képezi), és talál benne egy olyan rekordot, ami megfelel az általa keresett személy adatainak, hogyan lehet biztos abban, hogy ez az adat valóban hozzá tartozik? Könnyen lehet valaki más, akinek pont ugyanaz az irányító száma, születési dátuma, neme stb. Azonban az is sejthető, hogy ha a rekord egyre pontosabban megfelel a keresett személy adatainak, egyre valószínűbb, hogy ő lesz a keresett ember.

Azonban a Nature Communication-ben megjelent új tanulmány [9] szerint még az erősen mintavételezett anonimizált adathalmazokban is túl nagy az újra-azonosítás valószínűsége. A tanulmányban bemutatott statisztikai modell nagy pontossággal képes becslést adni egy adott populáció egyediségére. Egy populáció egy eleme akkor nevezhető egyedinek, ha az elem olyan jellemzők kombinációval rendelkezik, amely megkülönbözteti az adott populáció összes többi elemétől. A modellt demográfiai adatokon, és egészségügyi adatokból álló adathalmazokon tanították be. Bemutatták, hogy a modell erősen mintavételezett adathalmazon is nagy pontosságot ér el. Betanítás után a modell segítségével megmondható, ha van egy találatunk a mintavételezett adathalmazban, mekkora annak a valószínűsége, hogy a találat helyes, azaz megbecsülhető egy illető újra-azonosításának valószínűsége.



Ábra 5. Az a) diagrammon az átlagos egyediség eloszlása látható három (kék) és négy (narancs) demográfiai attribútum alapján. Szaggatott kék és narancs vonallal van jelölve William Weld kormányzó egyediségi indexe (ξ_x = 0.58, ξ_x = 0.997) rendre három és négy attribútum esetén. A b) diagrammon az átlagos egyediség van ábrázolva a demográfiai adatok számától függően. [9]

A modell működése szemléltethető egy példán keresztül. A modellt a Public Use Microdata Sample (PUMS) adathalmaz 5%-án tanították be [20], ami három adatot tartalmaz az amerikai lakosságról: ZIP kód, születési dátum és nem. A modell segítségével megbecsülhető egy illető egyedisége, illetve az alapján az újra-azonosíthatóság valószínűsége. Példaképpen vegyük William Weld kormányzót Latanya Sweeney tanulmányából, aki egy 1945. július 31-én született és Cambridge-ben élő férfi; a modell szerint 58% valószínűséggel egyedi a mintában [5. ábra (a): $\xi_x = 0.58$ és $\kappa_x = 0.77$]. ebből az következik, hogy ezen három adat alapján a de-anonimizálás sikerességének valószínűsége 77%. Azt is láthatjuk, hogy ha az adathalmaz tartalmazná még az illető gyerekeinek számát (ami William Weld esetén 5 gyerek), akkor az újra-azonosítás valószínűsége már 99,8% lenne (5. ábra (a) narancs vonal).

Az 5. ábra (b) részén látthatjuk, hogy az adathalmazban megjelenő attribútumok számának növekedésével a populáció egyedisége is gyorsan nő. 15 demográfiai adat esetén a Massachusetts lakosságára nézve az emberek 99.98%-a egyedinek számít. Az itt használt adathalmaz csak néhány attribútumot tartalmaz, de a korszerű adathalmazok ennél sokkal több attribútummal rendelkezhetnek. Például az Experian adatközvetítő által értékesített anonimizált adathalmaz 248 attribútumot tartalmaz háztartásonként [13].

Mint azt láthattuk, még az erősen mintavételezett adathalmazokban is nagy az újra-azonosítás valószínűsége, ami tovább nő az adathalmazban megjelenő attribútumok számával. Ezért ez a megfelelő anonimizás alkalmazásával küszöbölhető ki.

2.3.2 Anonimizálás nehézsége nagy adatban

Ahogy már arra kitértünk, a nagy adat anonimizálásának fő nehézségét az jelenti, hogy nem csak a rekordok (személyes profilok) száma nő meg jelentősen az adathalmazban, hanem az adatok attribútumainak száma is. Míg a korai esetekben az volt a jellemző, hogy az adatsorok számossága nagyságrendileg magasabb volt, mint az adathalmaz attribútumainak számossága, ez megváltozik; nagy adatban sokszor nagyságrendileg hasonló a két jellemző, sőt, a jellemzők száma akár meg is haladhatja az adatbázis méretét. Ez azt jelenti, hogy a néhány tíz jellemző helyett inkább több ezres jellemző lesz az adatban.

A megnövekedett dimenzionalitás eredményeképpen bár egyes jellemzők kihasználtsága gyakori lesz, a többséget csak néhány felhasználó fogja érinteni. Vegyünk

egy utalásokat tartalmazó adatbázist a példa kedvéért, amelyben a sorok az egyes felhasználók által elindított utalásokat jelzik (felhasználói rekordok, vegyesen természetes és jogi személyektől), az oszlopok az egyes utalásoknak a címzettjeit (ezek a jellemzők valójában szintén felhasználók). Az egyszerűség kedvéért hagyjuk figyelmen kívül a rendszeres jövedelem utalásokat. Míg ebben lesznek olyan címzettek, amelyek gyakran fogadnak be utalásokat (pl. online webshopok utalási fizetési lehetőséggel), a legtöbb felhasználó alig néhány utalást fog csak fogadni.

Ez a ritkasság nehézséget okoz valamennyi típusú anonimizálás esetén, ugyanis a ritkán használt jellemzők jó alapot adnak az azonosításhoz: ha tudjuk, hogy valaki mely ritka elemekkel hozható összefüggésbe (rétegfilmek, utalások magánszemélyeknek, ritka betegségek, stb), akkor ezek őt jól azonosítják. Márpedig ezek jó részével kezdeni kell valamit ahhoz, hogy az anonimizálás eredményes legyen.

Ez egyértelműen kihívást jelent a véletlenítésen alapuló anonimizálási eljárásoknál. Valamennyi ide vonatkozó eljárást nehezíti, hogy úgy kell módosítani az adatokat (zaj hozzáadással, permutációval vagy a differenciális adatvédelemmel), hogy az adjon hihető letagadhatóságot (hiszen teljes sorokat felülírni a használhatóság miatt nem lehet), mindamellett ne változtassa meg zavaró mértékben a felhasználás szempontjából releváns eloszlásokat. Ezt a nehezen garantálható kritériumot leginkább a differenciális adatvédelem tudja nyújtani, ami újra-azonosítási támadásokkal szemben is képes védelmet biztosítani, azonban magas szakértelem igénye és komplexitása miatt kevés helyen alkalmazzák, ráadásul olyan kötöttségei is lehetnek, amelyek gyakran nem felelnek meg bizonyos üzleti elvárásoknak.

Az általánosításon alapuló anonimizálási módszerek a kvázi-azonosítók értékeit írják felül egy kevésbé specifikus értékkel. A korábban bemutatott tipikus anonimizálási módszerek, mint a k-anonimitás feltételezik, hogy az adathalmazban viszonylag kevés kvázi-azonosító szerepel. Nagy dimenziós adatok esetén azonban nem használhatóak hatékonyan, mivel a sok jellemző miatt számos attribútum szolgálhat kvázi-azonosítóként: az anonimizálási technikák erősen függenek a térbeli lokalitás fogalmától, ami nagy dimenziós térben nehezen definiálható mivel az adatpontok közötti távolságok kevésbé térnek el egymástól [14]. Tehát abban az esetben, mikor az adat sok olyan attribútumot tartalmaz, amely kvázi-azonosítóknak tekinthető, akkor nehézkessé válik az adatok anonimizálása a nélkül, hogy túlzott mértékű információt vesztenénk.

Mindezekre a problémákra megoldást adhat az adatok szintetikus előállítása, akár az ide tartozó generatív gépi tanulási eljárások. Azonban ezek sem tökéletesek, és nem képesek a generálandó minden adat vetületét pontosan modellezni. Ebből fakadóan nincs általánosan ajánlható technika, hanem mindig a konkrét felhasználáshoz kell igazítani az alkalmazott eljárást.

2.4 De-anonimizálási eljárások főbb történelmi állomásai

Az első közismert de-anonimizálási eljárást Latanya Sweeney hajtotta végre 1997-ben, egy adott közigazgatási körzethez tartozó közalmazottak álnevesített egészségügyi adatain [5]. Bár helytelenül, de az általános szakmai vélekedés szerint akkoriban az álnevesített adatokat is anonimizáltnak számítottak. Ha csak önmagukban vesszük figyelembe ezeket az adatokat, akkor valóban anonimnek tekinthetők. Sweeney felismerte, hogy az alapvető demográfiai adatok egyedi módon azonosítják az adathalmazban szereplő személyeket, és ez megnyitja a lehetőséget az álnevesítés utólagos feloldására.

Sweeney úgy kalkulált, hogy a nem és az életkor, mindössze 56 940 lehetséges kombinációt ad (78 éves életkort feltételezve: $78 \cdot 365 \cdot 2 = 56\,940$), ami a vizsgált adatbázisban lévő többség számára egyedi azonítást jelent. Megvásárolta 20 dollárért ugyanannak a körzetnek a szavazói adatbázisát (kb. 25 ezer fő), ami szintén tartalmazta ugyanezeket a demográfiai adatokat, de a nevekkel együtt. Végül Sweeney a nem és életkor adatokat kombinálva sikeresen de-anonimizálta William Weld egészségügyi adatait, aki akkoriban a Massachusetts állam (USA) kormányzója volt.

Ennek a munkának a folytatásaként Sweeney az vizsgálta, hogy a teljes populációra kiterjeszthető-e ez a de-anonimizációs támadás. 2000-ben megjelent tanulmányában megmutatta, hogy az amerikai népesség 87%-át (216 millió főt a 248 millióból) egyedileg azonosítja az irányítószám, nem és születési dátum adatpontok együttese [4].

Ezek a felismerések elindították a táblázatos formátumú adatok anonimizálásának kutatását, kezdve a Latanya Sweeney által publikált k-anonimitási eljárással [6]. Ezek az anonimizálási eljárások nem alkalmazhatóak a ritkás adathalmazokra, mert az egyes jellemzőket nem lehet hatékonyan összevonni vagy elnyomni. Azonban egészen 2008-ig nem létezett hatékony de-anonimizálási eljárás ezekkel szemben.

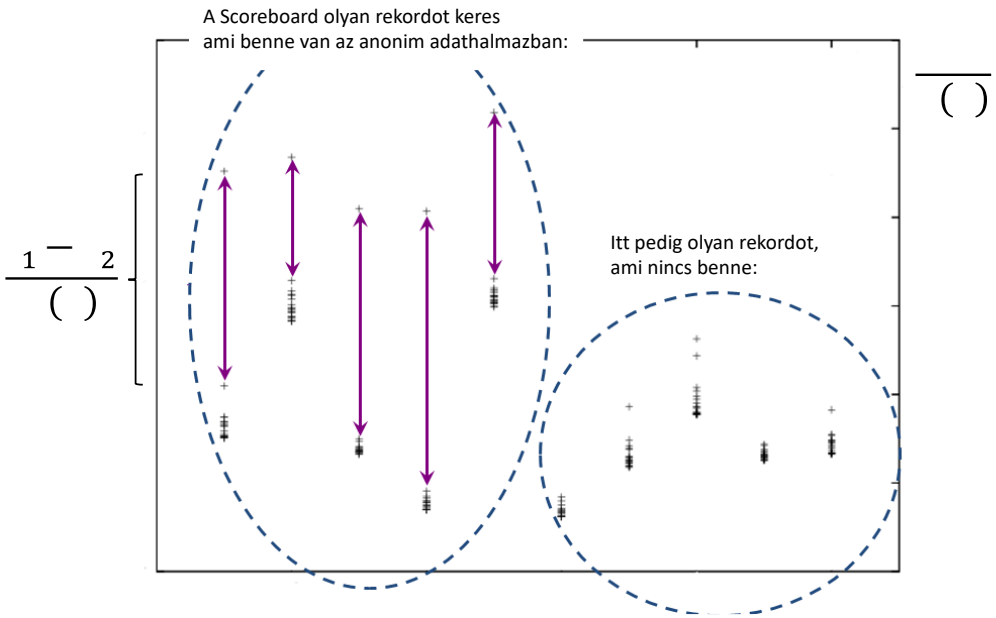
A Netflix 2008-ban online DVD kölcsönzéssel foglalkozott. Tevékenysége kulcskomponense a Cinematch elnevezésű ajánló algoritmus volt, amely a filmnézési előzmények alapján további filmeket kínál fel kölcsönzésre. Az ajánló algoritmus fejlesztése céljából a Netflix versenyt indított, s hogy a versenyzők tesztelhessék algoritmusaikat, elérhetővé tették 2006-ban 480 189 felhasználó álnevesített filmértékelést a 1999 december és 2005 december közötti időszakból [8]. Ez az adatbázis összesen 100m értékelést tartalmazott.

Narayanan és Shmatikov megmutatták, hogy akár egész pontatlan háttérismerettel is már de-anonimizálható a publikált felhasználói filmértékelések [8]. Az általuk készített Scoreboard nevű algoritmus anonimizált vagy álnevesített ($r' \in D'$) és ismert rekordok ($r \in D$) összehasonlításán és rangsorolásán alapszik. A pontozás során azokat a jellemzőket nagyobb hangsúllyal veszi figyelembe az algoritmus, amelyek általában véve kevésbé jellemzőek a teljes adathalmazra. A filmes témakörnél maradva ez könnyen érthető: a Scoreboard szerint jobban azonosít egy személyt, ha megnézett egy bizonyos szubkultúrának szóló filmet, mintha egy hollywoodi sikerfilmet.

Következő lépésben az (r', r) rekordpárok pontozása alapján megtörténik az azonosítatlan személyhez kapcsolódó rekordhoz potenciálisan tartozó identitások rangsorolása. Ezt jelöljük S-sel, amelyre $\forall s_i \in S: s_1 \geq s_2 \geq \cdots \geq s_n$. Ezen a ponton a Scoreboard megvizsgálja, hogy a legjobb találat elég jó-e, azaz csak akkor fogadja el, ha lényegesen jobb, mint a többi. Ezzel szorítja le az algoritmus a hamis pozitív találatok számát. Ezt a kulcsfontosságú kritériumot a szerzők kiugróság (angolul eccentricity) vizsgálatnak nevezik, amelyet a következőképp definiáltak:

$$\frac{s_1 - s_2}{\sigma(S)} > \theta,$$

ahol θ egy futásidőben konstans paraméter, amely az algoritmus mohóságát szabályozza. A kiugróság vizsgálat elve univerzális és hasznos eleme a de-anonimizálásnak, olyannyira, hogy más tanulmányok is beépítettek hasonló mechanizmusokat, mint például a későbbiekben tárgyalt gráf de-anonimizációs algoritmus [2].



Ábra 6. Magyarázat a kiugróság mérésének jelentőségére a Netflix által publikált adatokon. Az x tengelyen különböző ismert rekordok helyezkednek el, az y tengelyen pedig az ezekhez tartozó hasonlósági mérések eredményei az anonimizált adathalmaz rekordjaihoz képest. Az ábra jól szemlélteti, hogy az anonimizált adatbázisban is szereplő rekordok kiugrósága jelentősen eltér azokétól, amelyek abban nem szerepelnek. (az ábra forrása: [16])

A [8] tanulmány szerzői a Scoreboard algoritmust először egy-egy véletlenszerűen kiválasztott felhasználó megtalálásával tesztelték a Netflix adathalmazban. A keresett felhasználó adatait az IMDb adathalmazból választották, ami a támadó de-anonimizáláshoz használt háttértudásának felelt meg ebben az esetben. Bár az IMDb felhasználók többsége 20-nál több értékelést írt már, a támadás ellenőrzésének életszerűségét demonstrálandó csupán 2-8 értékelést választottak ki a Scoreboard számára, mint háttérinformációt. Ezekből az eredményekből emelünk ki a következő bekezdésben néhányat.

Az algoritmus több mint 80%-os találati arányt ért el mindössze 6 értékelés figyelembevétele alapján, ha a film pontozása pontosan ismert volt (1-5 csillag), de az értékelés dátuma csak ±14 nap pontossággal. Ez akkor is működött, ha a 6 értékelésből egy hibás volt. Ugyanez az eljárás 8-ból 7 helyes értékelés esetén a de-anonimizálás találati aránya 90% felé nőtt. A szerzők azt is megmutatták, hogy a kiegészítő információ pontatlanságát jól lehet ellensúlyozni az értékelések számával.

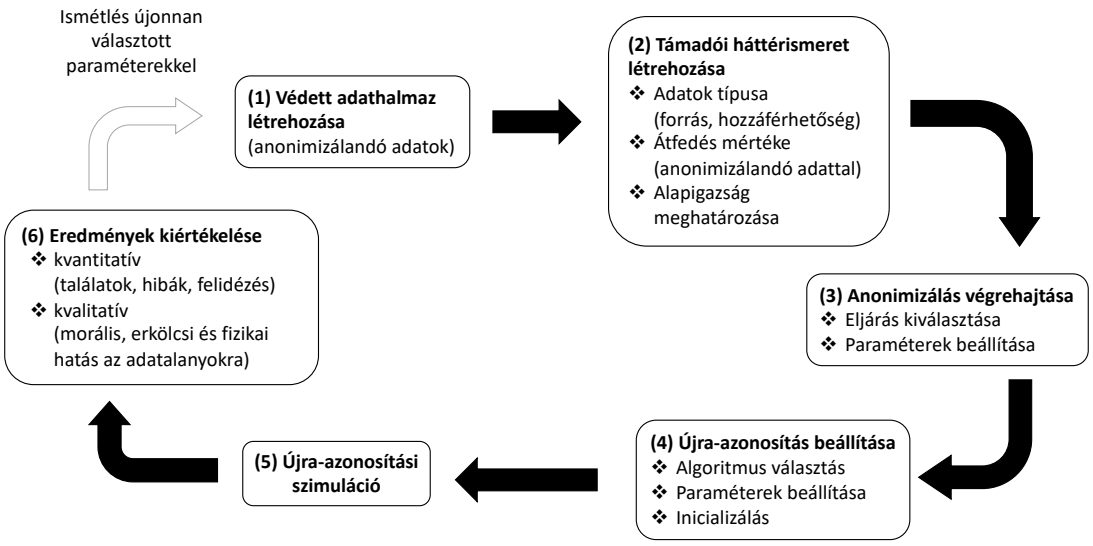
Mivel ezek a támadások sikeresek voltak, a következő kísérletben arra voltak kíváncsiak, hogy ha ezeket a rekordokat törlik az anonimizált adatok közül, akkor az algoritmus képes lesz-e ezt jelezni. Ez a Scoreboard esetén azt jelenti, hogy a kiugróságnak a küszöbérték alatt kell maradnia, mert egy találat sem lehet kimagaslóan releváns. Az előző kísérletekben a hiányzó rekordok téves de-anonimizálását is el tudta kerülni a Scoreboard algoritmus; nem befolyásolta a találati arányt, hogy ha a keresett felhasználók között vegyesen voltak olyanok, amelyek szerepeltek az anonim adatok között, és olyanok is, amelyek nem.

Narayanan és Shmatikov azt is megmutatták, hogy az adatok enyhe módosítása nem csökkenti hatékonyan a de-anonimizálás kockázatát. A Netflix által kiadott adathalmaz ilyen volt, és ez nem akadályozta meg a Scoreboard algoritmust: kettő, az anonimizált adatban szereplő ismerősüknél 1/306 és 5/229 értékelés tért el az eredetitől, így őket is hatékonyan megtalálta az eljárás. Ebből az következik, hogy akár egy rövidebb beszélgetés vagy néhány értékelés az IMDb-n visszakereshetővé teszi azokat, akiknek értékeléseit a Netflix publikálta.

A Scoreboard algoritmus egyik fő újdonsága az adott alkalmazásban nyújtott hatékonyságban és sikerében rejlett, hiszen megmutatta egy álnevesített és anonimnek tekintett adathalmazról, hogy az valójában újra felruházható az eredeti felhasználói identitásokkal. Azonban ennél fontosabb eredménye, hogy az első algoritmus volt, ami megmutatta, hogy a mintavételezett és enyhén zajosított nagyméretű, ritka adatbázisok is de-anonimizálhatóak. Végezetül pedig a bemutatott Scoreboard algoritmus kellően rugalmas ahhoz, hogy más típusú adaton is alkalmazható legyen, így alkalmas arra, hogy későbbi újraazonosítási eljárásokhoz sablonként szolgáljon.

3 Anonimizálás ellenőrzése és a GDPR

Az anonimizálás hatékonysága és egyben használatának jogalapja azon múlik, hogy mennyire képes ellenállni az újra-azonosítási kísérleteknek. Továbbá a GDPR 26. cikkje [1] több ponton is meghatároz olyan követelményeket, amelyeket anonimizálásnál figyelembe kell venni. Nézzük meg, hogy ezek figyelembevétele mellett hogyan lehet az anonimizálás hatékonyságát értékelni. A későbbiekben egy erre épülő módszertant fogunk követni tranzakció anonimizálás ellenőrzésére.



Ábra 7. Anonimizálási eljárások ellenőrzésének folyamata.

Az első lépésben (7. ábra, (1)-es lépés) meg kell határozni, milyen adatot akarunk anonimizálni. Az adatból el kell távolítani a természetes személyt egyértelműen azonosító adatokat, ugyanis ez az anonimizálás előfeltétele. A 26. cikk szerint az számít csak anonim adatnak, ami nem vonatkoztatható természetes személyre, vagy olyan anonimizálási eljárás alá vetett személyes adat, amely többé nem hozható kapcsolatba az adatalannyal. Ez utóbbinál a GDPR azt is elfogadhatónak tartja, ha elvben lehetséges volna az anonimizálás feloldása, de a gyakorlatban a megvalósítása nem praktikus, például lehet, hogy túl sok, nehezen hozzáférhető adatra van szükség, vagy a folyamat szakértelem igénye, komplexitása indokolja.

Mielőtt azonban az anonimizált adatokat publikáljuk, az alkalmazott anonimzálási eljárást több szempont alapján meg kell meg kell vizsgálni. Ezek szerepelnek a 7. ábra (2-6) lépéseiben. A GDPR előírja, hogy egy természetes személy „*azonosíthatóságának meghatározásakor minden olyan módszert figyelembe kell venni* [...]*, amelyről észszerűen feltételezhető, hogy az adatkezelő vagy más személy a természetes személy* [...] *azonosítására felhasználhatja*" [1]. Ez több releváns szempontot magába foglal, a lehetséges választásokat és azok kombinációit vizsgálni kell.

Először is, a támadás sikerét jelentősen meghatározó tényező, hogy mit feltételezünk a rosszindulatú fél rendelkezésére álló információkról (7. ábra, (2)-es lépés). Ezt hívjuk háttérinformációnak. Például hatékonyabb lehet egy támadás, ha a támadó fél ugyanolyan típusú háttérinformációval rendelkezik, mint az anonimizált adat, azaz például személyes azonosítókat tartalmazó tranzakciós adatbázis áll rendelkezésére. A hálózatokká alakított tranzakciós adatbázisok újra-azonosításához kommunikációs vagy közösségi hálózatokból származó információkat is fel lehet használni. Sokat számít, hogy ezek a de-anonimizálással támadott adatbázisokkal hasonló méretűek-e, illetve a háttérinformációban hány személy található meg az anonim adatokban lévő természetes személyek közül, azaz mekkora az átfedés mértéke (és minősége).

Az újra-azonosításnak ezen paramétereinek eltérő költség és időigénye lehet, és korlátozza a lehetséges támadók személyét. Például bizonyos adatokhoz csak egyes telekommunikációs operátorok vagy közösségi hálózatok üzemeltetői férhetnek hozzá. Ezt azonban szintén szükséges vizsgálni, esetleg további szempontokkal, mint a támadás végrehajtásához szükséges szakértői tudás. A 26. cikk alapján elvárt, hogy az „*összes objektív tényezőt figyelembe kell venni, így például az azonosítás költségeit és időigényét, számításba véve az adatkezeléskor rendelkezésre álló technológiákat, és a technológia fejlődését*" [1].

Bár az idézet megfogalmazásából egyértelműen nem következik, de a vizsgálat tárgya alá értendők azok a technológiák is, amelyek adatforrásként szolgálhatnak. Például lehetséges kapcsolatrendszert felépíteni az együtt töltött idő alapján, de ennek végrehajtása komplex és komoly szakértelmet igénylő feladat [11].

Ha már tudjuk a fenti paramétereket, amelyek a támadó fél erősségét határozzák meg, lehetőségünk nyílik a de-anonimizációs eljárások gyakorlati vizsgálatára. Bizonyos esetekben az elméleti vizsgálat lehetősége is adott lehet, ez azonban az adatok

mennyisége és a probléma komplexitása miatt a legtöbb esetben praktikusan nem kivitelezhető.

A gyakorlati vizsgálat során az adott feltételek szerinti támadás szimulációját lefuttatjuk, és megvizsgáljuk annak végeredményét. Az értékeléshez alapul azok a rekordok szolgálnak, amelyek mindkét adatbázisban jelen vannak; s mivel szimuláltuk a támadást, ezért tisztában vagyunk ezeknek a számával, és az anonimizált és háttérbeli adatok közötti összerendelésekkel is (ez az ún. alapigazság, angolul ground truth).

Az anonimizálás vizsgálása során érdemes egy adott eljárást használni fix paraméterekkel, hiszen így többféle erősségű és jellegű támadó esetén átfogóbb képet kapunk az eljárás hatékonyságáról. Szúrópróba-szerű ellenőrzésnél itt is meghatározhatjuk az anonimizálás algoritmusát és paramétereit. Majd végrehajtjuk az anonimizálást a védendő adatbázison (7. ábra, (3)-as lépés).

Majd választani szükséges a rendelkezésre álló de-anonimizálási algoritmusokból, azoknak a lehetséges paramétereit és inicializálási módját ki kell választani, majd ennek megfelelően a kezdő anonim-háttérismeret rekord párosításokat meg kell adni (inicializálás). Ezt követően lefuttatjuk az újra-azonosítási eljárást. Ezek a 7. ábrán lévő módszertanban a (4-5)-ös lépések.

A de-anonimizációs algoritmusok az alapigazsághoz hasonló párosításokat adnak meg, amelyekben az általuk helyesnek vélt adatpontokat kötik össze az anonimizált adathalmazban és a háttérismeretben. A kiértékelés során ezeket a párosításokat kell összevetni az alapigazsággal (7. ábra, (6)-os lépés), amelyek így lehetnek helyesek (ún. true positive, röviden TP), illetve tévesek (ún. false positive, röviden FP). A meg nem talált párosítások arányának mérésére a felidézést érdemes használni (ún. recall). Ezekkel a metrikák gyakorlati alkalmazásával foglalkozunk a következő fejezetben.

Ezeket a metrikákat figyelembe véve, alapvetően kétféle típusú újra-azonosítási algoritmus létezik, a mohó és diszkrét [10]. A mohó algoritmusok jellemzően magas felidézést képesek elérni, azonban a hibázási arányuk viszonylag magas, jellemzően 10-50% körül mozog. Ezzel szemben a diszkrét algoritmusoknak a felidézése mérsékeltebb, de a hibák aránya tipikusan 10% alatt mozog. Általánosságban elmondható, hogy a mohó algoritmusok a gyakorlatban nem jól használhatóak, mert túl zajos eredményt

produkálnak. Ez nem meglepő, hiszen igen korlátozottan használható egy találati lista, ha tudjuk, hogy az eredmények kb. harmada-fele téves.

A támadás szimulációnak lefuttatása utáni feladat kiértékelni, hogy a támadás az adott feltételek mellett milyen arányban volt képes az anonimizált adatból azonosíthatót helyreállítani. A támadás sikerére vonatkozóan a GDPR nem határoz meg konkrét kritériumot, miközben az is nyilvánvaló, hogy tökéletes anonimizálás nincs; így az anonimizálási eljárás hatásosságának mérlegelése a mi szubjektív feladatunk marad. Ennek alapja lehet, hogy anonimizálás után hány személyt voltunk képesek az elvi maximumból megtalálni (felidézés), illetve az algoritmus által visszaadott hiba arány elfogadható-e.

Összhangban az adatvédelmi hatásvizsgálattal (angolul az ún. Data Protection Impact Assessment, DPIA), amelyet a GDPR megkövetelhet bizonyos esetekben a támadások kockázatának kiértékelésére, érdemes figyelembe venni, hogy a sikeres támadásban érintett személyeken kívül milyen hatása van a támadásnak az adatalanyokra. Ugyanis több mindentől függ a támadás következményeinek súlyossága, például attól, hogy a de-anonimizálás során a teljes érzékeny információhalmazhoz hozzáfér-e a támadó, vagy csak egy részéhez. Az is lehet, hogy a támadás során megszerzett védett információk kiszivárgásának kockázata csak mérsékelt.

A fenti vizsgálatot valamennyi ésszerűen feltételezhető támadóra el kell végezni (26. cikk, GDPR [1]), és a de-anonimizálás kockázatát ezeknek összességében kell vizsgálni. Amennyiben kedvezőek az eredmények, az anonimizálás eszközét felhasználhatjuk arra, hogy az adatokat kivonjuk a GDPR hatálya alól, ugyanis a 26. cikk megállapítja [1], hogy az „*adatvédelem elveit* [...] *anonim információkra nem kell alkalmazni*”.

4 Tranzakciós adatok de-anonimizálása

A következőkben a tranzakciós adatok de-anonimizálását mutatjuk be úgy, hogy a tranzakciós adatok gráf struktúrára van visszavezetve (amelyre anonim, célzott adatként hivatkozunk később). Feltételezzük, hogy a támadó fél rendelkezésére áll ehhez egy másik gráf formátumú adatbázis (háttérismeret, kiegészítő adat).

4.1 Az adatok hasznosítása

Anonimizálás esetén két fő szempontot kell figyelembe venni a megfelelő algoritmus kiválasztásához: az adat hasznosíthatóságának mértékét és az anonimizálás erősségét. Az adatok hasznosíthatóságára nehéz általános követelményt megfogalmazni, hiszen ez nagyban függ az alkalmazástól, és alkalmazásonként eltérő lehet.

Tranzakciós adatok esetében ez lehet olyan információ, ami elsősorban a csomópontokhoz kapcsolódik, például, hogy az egyes szereplők milyen anyagi háttérrel és költési hajlandósággal rendelkeznek. Kapcsolódhat az élekhez, ami például vonatkozhat a tranzakciók tárgyára, összegére stb. Illetve lehet, olyan, ami a hálózat struktúrája szempontjából érdekes, például milyen jellegű pénzmozgási minták ismerhetők fel a hálózaton belül. Lehetnek egyszerű metrikák, lehet hasznos információ a fokszám eloszlás, vagy valamilyen központiság metrika. (A későbbiekben ilyen általánosan ismert és elfogadott gráf struktúrához kötődő metrikákat fogunk alkalmazni.)

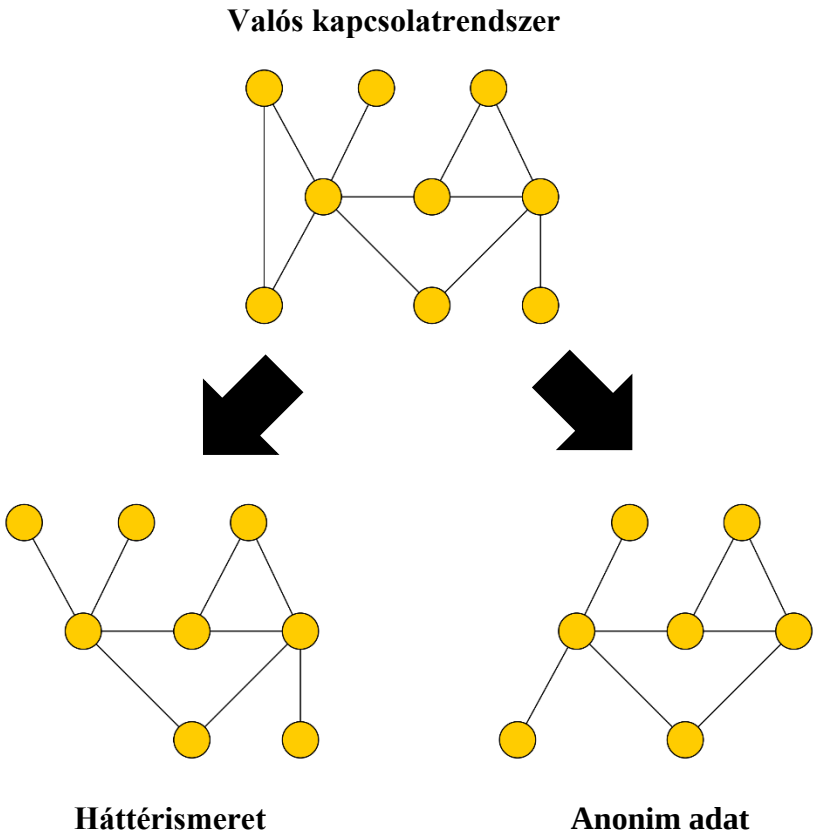
4.2 Adatok előkészítése

Általában nem áll rendelkezésre megfelelő háttérismeret, amely az anonimizálandó adathoz úgy illeszkedik, hogy a két adatbázison tesztelni lehessen anonimizálási, majd de-anonimizálási eljárásokat. Emögött általában egyszerű okok állnak a hátterében: nem lehet megfelelő háttérismeret adatbázishoz jutni, vagy ha mégis sikerül, nem tudjuk mesterségesen helyreállítani az alapigazságot (csomópontok összepárosítása a két adatbázisban annak megfelelően, hogy melyik csomópontok felelnek meg ugyanannak a személynek). Az ilyen esetekben először szintetikusan kell előállítanunk a támadó szimulált háttérismeretét az anonimizált adatok mellé.

4.2.1 Adatbázis-párok szintetikus előállítása

A legegyszerűbb eljárás, ha egyszerűen csak lemásoljuk az anonimizált adatokat. Azonban ez igen erős támadót feltételezne, és nem ad lehetőséget többféle, életszerűbb támadó vizsgálatára.

A szakirodalom többféle eljárással találkozhatunk, amelyek két fő csoportra bonthatók: zaj hozzáadásra és mintavételezésre épülő eljárások. Az előbbiek fő problémája, hogy nehéz a zaj eloszlását úgy beállítani, hogy a végeredményképp létrejövő hálózat hihetően életszerű legyen. Ezért célszerűbb a második kategóriába tartozó eljárásokat alkalmazni, melyek jobban tükrözik a valóságban jelenlévő folyamatokat: az egyes természetes személyek közötti kapcsolatrendszer sohasem egy-az-egyben képződik le egy szolgáltatásban megjelenő kapcsolatrendszerre, hanem mindig csak annak egy része (ld. 8. ábra).



Ábra 8. Amennyiben nem áll rendelkezésre megfelelő háttérismeret adatbázis, szintetikus módon kell generálni háttérismeret és anonim adatbázis párt, amelyen az anonimizálás, és majd a de-anonimizálás eljárásokat tesztelhetjük.

A szakirodalomban két szintetikus adat előállítási eljárás terjedt el széles körben. Mindkettőnél kiindulásként rendelkezésre áll egy G gráf, amely a valós kapcsolatrendszert szimbolizálja, és ebből származtatjuk a háttérismeret és az anonim adatot is.

Az első eljárást nevezzük *NarPert*nek [2]. Ennek paramétereként meg kell adni, hogy a háttérismeret és anonimizálandó gráfok mennyire legyenek hasonlóak. A csomópontok halmazainak hasonlóságát jelöljük α_v-vel, az élek halmazainak hasonlóságát pedig α_e-vel. Ez az eljárás a Jaccard hasonlóságot használja a generált adathalmazok hasonlóságának mérésére:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Ennek előnye az egyszerű metszet számításhoz képest, hogy képes figyelembe venni, ha az egyes adathalmazok mérete jelentősen eltér egymáshoz képest. Az eljárás először lemásolja a G gráfot két példányban, majd először csomópontokat töröl mindkét példányban (független módon) a kívánt α_v eléréséig. Majd ugyanígy egymástól függetlenül elkezd éleket törölni a példányokban α_e eléréséig.

A másik eljárás, amelyet nevezzünk *SamplePert*-nek, az élek mintavételezésére épít, és úgy másolja le a G gráfot, hogy az éleinek csak egy meghatározott hányadát (pl. 70%) másolja át [22][23][24]. Ennek a megoldásnak a hátránya, hogy korlátozottan kontrollálható mértékben tér el csomópontok halmaza a háttérismeret és az anonim adat között, kevésbé finomhangolható, mint a NarPert.

Például a 60%-os mintavételezési arány jellemzően alsó határként szerepel szimulációkban (pl. lásd a [22] művet), ez azonban a csomópontok átfedése szempontjából még mindig erős támadó modell a NarPert eljárás szempontjából, ugyanis ez $\alpha_v = 0{,}77$ és $\alpha_e = 0{,}42$ értékeknek felel meg. Ennek, mint alsó határnak, nem életszerű mivoltát jól mutatja a [2]-ban közölt, valódi adatokkal dolgozó támadás, ahol az átfedés mindössze 27 ezer csomópont volt, ami az eredeti cikkben közölt adatok alapján kb. $\alpha_v = 0{,}0077$ értéknek felel meg. Ezt a tulajdonságot a SamplePert eljárás nem tudja kontrollálni.

4.2.2 Anonimizálási eljárások

A következőkben áttekintünk néhány gráf struktúrájú adatokra alkalmazható anonimizálási eljárást. Ezeket, hasonlóan az általános anonimizálási technikákhoz, be tudjuk sorolni a korábban ismertetett iskolák valamelyikébe (ezt is megadjuk).

Az egyik legegyszerűbb eljárás a véletlenítés iskolájába tartozó, permutációs technika a *Switch(k)*: a k paraméternek megfelelő százaléknyi élt áthelyezünk a gráfon belül [25]. Bár az adott hálózat méretétől is függhet a k megfelelő értéke, jellemzően 5-10% körül alkalmazzák az eljárást a szakirodalomban. Ez akár több tízezer – 1 millió él áthelyezését is jelenti. Előnye az egyszerűsége és sebessége, hátránya, hogy a módosítás nem követi a gráf eredeti strukturális eloszlásait.

A *k-DA* (angolul k-degree anonymity, magyarul k-fokszámú anonimitás) eljárás az általánosítás iskolájába tartozik, azon belül is a k-anonimitás csoportba [26]. Ez az eljárás garantálja a k-anonimitás tulajdonság megvalósulását minden csomópontra azok fokszáma alapján. Ezt úgy éri el, hogy a gráf csomópontjait és éleit megőrzi, de új éleket generál hozzá úgy, hogy az eredeti fokszám eloszlástól a legkevésbé eltérő, de a k-anonimitásnak megfelelő fokszámolást alkalmaz. Az új élek felvétele jelentősen módosíthatja a gráf strukturális jellemzőit, de ez is egy koncepcionálisan egyszerű anonimizálási eljárás.

A Pygmalion nevű (*DP(ε)*-nek jelöljük), differenciális adatvédelemre épülő eljárás szintetikus gráfot hoz létre az eredeti alapján [27]. Az eredeti gráf (szomszédos) csomópont párjainak fokszámainak együttes eloszlását veszi alapul, majd ehhez a differenciális adatvédelemben használt Laplace zajt ad hozzá. Ez után generálja le a szintetikus gráfot, ügyelve, hogy a generált gráf nagyban hasonlítson az eredetire. Alapvetően jól követi az eredeti eloszlásokat, de nem mindenben; például a globális klaszterezési együtthatót a másik két módszernél rosszabbul képes tartani.

4.3 De-anonimizálási algoritmusok

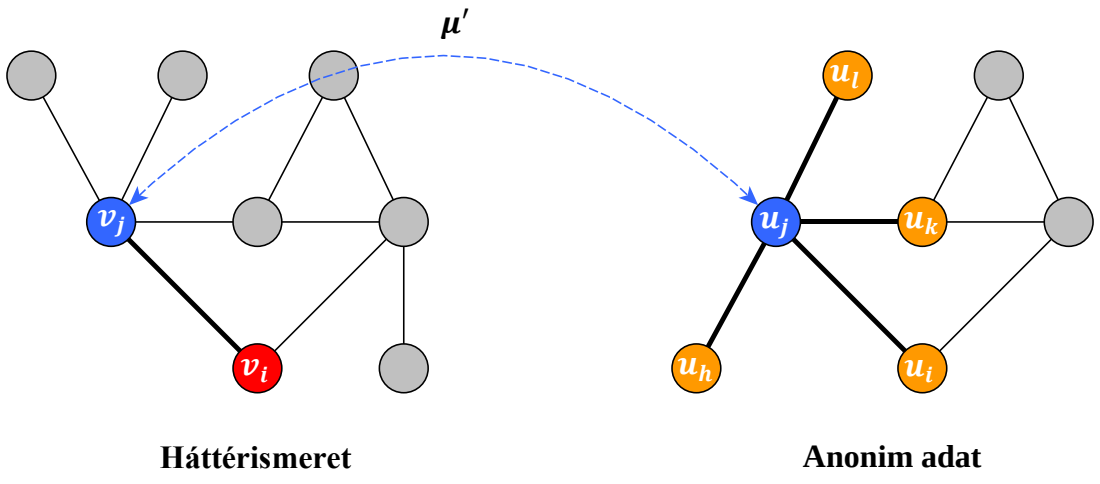
Ahogy a bevezetőben is bemutattuk, a gráf de-anonimizációs algoritmusok alapvetően két fázisból állnak: egy inicializálási fázisból, és egy ezt követő, iteratív alkalmazású terjedési fázisból. Bizonyos algoritmusok csak terjedési fázissal rendelkeznek, ezek önindító módon képesek elindulni. Jelen tanulmányban azonban csak olyan algoritmusokat vizsgálunk, amelyek mindkét fázisra építenek (mivel ezek az

algoritmusok tekinthetők a leghatékonyabbnak). Ez nem jelent az eredmények szempontjából korlátozást, mivel ezeket az eljárásokat a leghatékonyabb algoritmusok között tartja számon a szakirodalom.

4.3.1 Általános jellemzők

A továbbiakban az egyszerűség kedvéért jelöljük a háttérismeretet, azaz az ismert identitásokkal rendelkező gráfot G_I-nek (V_I csomópontokkal, és E_I élekkel), az anonim adatot pedig hasonlóan G_A-nak (V_A, E_A). A V_I és V_A között létezik egy μ hozzárendelés, amit nevezzük az alapigazságnak, és ami meghatározza, hogy az egyes azonosítatlan csomópontok (V_A) mely azonosított identitással rendelkező személyeknek felelnek meg (V_I), azaz $\mu : V_I \rightarrow V_A$. Minden de-anonimizálási algoritmus célja egy olyan μ' hozzárendelést megtalálni, ami a lehető legközelebb van az alapigazsághoz.

Az inicializálás során létrehozunk egy kezdeti μ_0 hozzárendelést. A terjedési fázis elkezdi kiterjeszteni a μ_0 hozzárendelést, új kapcsolatokat vesz fel, esetleg felülvizsgálja a már meglévőket. Egy általános alapelvre épít itt valamennyi algoritmus: ha létezik egy pár szomszédos csúcs az egyik hálózatban, akkor azok nagy valószínűséggel szomszédosak lesznek a másikban is (feltéve, hogy mindkettő létezik benne). Ezt az algoritmusok úgy használják ki (ld. 9. ábra), hogy ha keressük egy csúcshoz a megfelelő hozzárendelést (piros csúcs; v_i), akkor a vele szomszédos, már hozzárendeléssel bíró csúcsok (bal kék csúcs; v_j) hozzárendeltjének (jobb kék csúcs; u_j) a szomszédai között (narancs csúcsok; u_h, u_i, u_k, u_l) keressük a számára megfelelő hozzárendelést.



Ábra 9. Alapelv a meglévő hozzárendelések felhasználására az újraazonosítás folyamatában. Ha keressük egy csúcshoz a megfelelő hozzárendelést (piros), akkor a vele szomszédos, már hozzárendeléssel bíró csúcsok (bal kék) hozzárendeltjének (jobb kék) a szomszédai között (narancs) keressünk.

Az algoritmus által előállított μ' hozzárendelések halmaza a végeredmény, ami alapján értékelhető a működése. Ezt összevetve a μ alapigazsággal, meghatározható a felidézés (az alapigazságban szereplő és megtalált csomópontok számának aránya az alapigazság teljes méretéhez) és a téves találatok aránya (hány hozzárendelés volt téves). Ezzel a két metrikával meghatározható, hogy egy de-anonimizációs algoritmus eredménye hogyan viszonyul az adathalmaz egészéhez, illetve az általa megadott eredmények mennyire pontosak.

4.3.2 Nar: a de-anonimizáló eljárások prototípusa

Az első pontos, nagyléptékű **de-anonimizálásra alkalmas algoritmust Narayanan és Shmatikov** tette közzé 2009-ben [2], melyre a továbbiakban *Nar* elnevezéssel hivatkozunk. Eredeti munkájukban az inicializáláshoz 4-es klikk párok közötti hozzárendelést kerestek, de ez többféle egyéb metrikára lecserélhető, mint például a legmagasabb fokszámú csúcsok vagy valamely központiság metrika szerinti egyezés alapján történő hozzárendelésre [28].

A terjedési fázisának működése egyszerű, egy iteráció a következő lépésekből áll:

1. Válasszunk ki egy csúcsot, amelyhez még nem létezik hozzárendelés. Ezt a keresést addig ismételjük, amíg van szabad csúcs.
2. Keressük meg az ehhez a csúcshoz tartozó lehetséges hozzárendelteket; ezek lesznek a jelöltek.
3. Pontozzuk és rangsoroljuk a jelölteket.
4. Ha van a jelöltek között kiugróan magas pontszámú, akkor hozzunk létre hozzárendelést erre a csúcsra.

A következőkben a működését a 9. ábrán látható jelölések alapján szemléltetjük. Tegyük fel, hogy már létezik egy hozzárendelés v_j és u_j között. Ekkor válasszuk ki v_j egy hozzárendeléssel nem rendelkező szomszédját, v_i-t (1. lépés). Ahogy létezik él v_i és

v_j között, feltételezhetően létezik él u_i és u_j között is; ezért a jelöltek az u_h, u_i, u_k, u_l csúcsok lesznek (2. lépés).

A *Nar* algoritmus a jelölteket a koszinusz hasonlóság alapján pontozza:

$$\mathrm{CosSim}(v_i, u_j) = \frac{|\text{közös hozzárendelések}|}{\sqrt{\deg(v_i)} \cdot \sqrt{\deg(u_j)}}$$

Azonban mivel a $\sqrt{\deg(v_i)}$ azonosan szerepel valamennyi esetben (konstans), ezt egyszerűen elhagyjuk:

$$\mathrm{NarSim}(v_i, u_j) = \mathrm{s}_j = \frac{|\text{közös hozzárendelések}|}{\sqrt{\deg(u_j)}}$$

Ezután a jelöltek a hasonlóságuk alapján rangsorolhatók (3. lépés). Hogy a legjobb jelölt megfelelő-e, azt ugyanazon a kiugróság metrikával mérjük, mint amit a Netflix esetén bemutatott de-anonimizációs eljárásnál is bemutattunk. A jelöltek pontszámainak halmazát jelölje S, melyeket csökkenő sorrendbe rendezünk, úgy, hogy s_1 értéke a legnagyobb. Ekkor az algoritmus θ paramétere alapján akkor veszünk fel új hozzárendelést a vizsgált v_i, u_j csúcsok között, ha az u_j pontszáma a legmagasabb, és megfelel az alábbi feltétlenek a második legjobb jelölthöz képest (4. lépés):

$$\frac{s_1 - s_2}{\sigma(S)} > \theta,$$

ahol a $\sigma(S)$ az adott halmaz szórását jelöli. Ezen feltétellel szabályozható az algoritmus mohósága. A kisebb θ értékeknél az algoritmus könnyebben elfogad új hozzárendeléseket, míg nagyobb értékek konzervatív működést eredményeznek. Ezen keresztül végső soron az algoritmus által megvalósított találati- és hiba arányok közötti trade-off válik szabályozhatóvá.

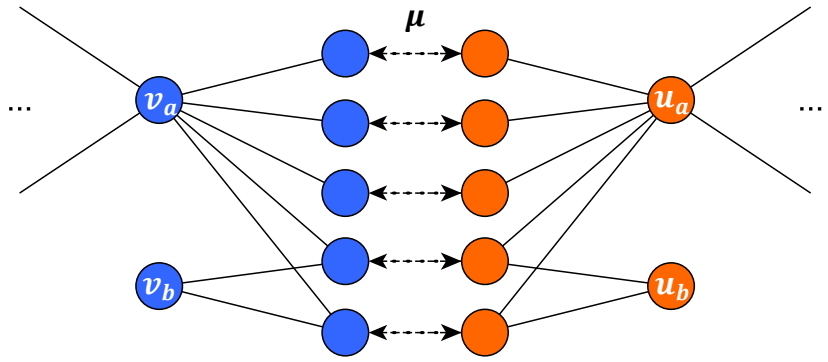
Az algoritmus hatékonyságát a szerzők valós közösségi hálózatokból gyűjtött adatokon demonstrálták [2]: gyűjtöttek 224 ezer felhasználót a Twitter közösségi hálózatból (amit később az „anonim” adatnak tekintettek), illetve 3,3 millió felhasználót a Flickr képmegosztó oldalról. Az alapigazság felállításához először meghatározták a profilokban szereplő felhasználónév, név és hely információk alapján azokat a felhasználókat, akik mind a két hálózatban szerepeltek; ezek alapján összesen 27 ezer felhasználói profilt

sikerült azonosítani. Ezeknek helyességét szúrópróbás szemrevételezéssel megerősítették.

A demonstrációhoz az inicializálási fázishoz véletlenszerűen kiválasztottak 150 népszerű felhasználót (az alapigazságból), akikkel szemben annyi elvárásuk volt, hogy legalább 80 kapcsolattal rendelkezzenek. A terjedési fázis lefuttatása után az alapigazságban szereplő felhasználók 30,8%-át sikerült helyesen megtalálni, 12,1%-át helytelenül. A fennmaradókhoz (57,1%) az algoritmus nem adott meg semmilyen hozzárendelést. Ezek az eredmények jól tükrözik az algoritmus hatékonyságát: két méretben és jellegben eltérő hálózat alapján képes volt azonosítani az átfedés egy jelentős részét.

4.3.3 További eljárások

A Bumblebee algoritmus. Noha a Nar algoritmus volt az a műfajteremtő algoritmus, amely először volt képes nagy méretű kapcsolati adatokat de-anonimizálni, voltak bizonyos hibái. Például döntési helyzetekben torzít, ha a jelöltek között jelentős a fokszám eltérés. Vegyük példának a 10. ábrán látható helyzetet, ahol a v_a és v_b hozzárendelési párjait keresi. Tegyük fel, hogy $\deg(v_a) = \deg(u_a) = 100$ és $\deg(v_b) = \deg(u_b) = 2$, és van ezek között már 5 hozzárendelés, melyek az ábrán látható módon átfedésben vannak.



Ábra 10. Döntési helyzet, melyben a Nar döntési mechanizmusa rossz döntést hoz.

Ebben a helyzetben a Nar algoritmus az alábbi táblázatban szereplő hasonlóságokat fogja kiszámolni.

	v_a	v_b
u_a	$\frac{5}{\sqrt{100}}$	$\frac{2}{\sqrt{2}}$
u_b	$\frac{2}{\sqrt{100}}$	$\frac{2}{\sqrt{2}}$

Táblázat 1. Nar pontszámok hozzárendelések kiszámításához.

Mivel a rangsorolásnál a nagyobb érték előnyben részesül, ezért az ehhez hasonló helyzetekben a Nar nem tudja megfelelően hozzárendelni a nagyobb fokszámú csomópont megfelelőjét, mivel a kisebb fokszámút előnyben fogja részesíteni. Ezt küszöböli ki a Bumblebee nevű algoritmus [10], amely működési vázát tekintve megegyezik a Nar algoritmussal, kivéve a hasonlósági metrikát. A továbbiakban az algoritmusra Blb rövidítéssel hivatkozunk, amely az alábbi metrikát használja:

$$\mathrm{BlbSim}(v_i, u_j) = \mathrm{s}_j = |\text{közös hozzárendelések}| \cdot \left(\min \left(\frac{\deg(v_i)}{\deg(u_j)}, \frac{\deg(u_j)}{\deg(v_i)} \right) \right)^{\delta}$$

Ez a képlet szimmetrikusan képes kezelni, ha a fokszámok jelentősen eltérnek; minél nagyobb az eltérés, annál kisebb az esély a két csomópont közötti hozzárendelés létrehozására. A fenti képletben a δ paraméterrel szabályozható, hogy az algoritmus a csomópontok fokszám eltérését mekkora mértékben vegye a figyelembe. A mérési tapasztalatok alapján ezzel a paraméterrel szintén az algoritmus mohóságát lehet állítani, hasonlóan a θ paraméterhez. Kiszámítva az előző példában megadott döntési helyzetre a Blb algoritmus értékeit ($\delta = 0{,}5$) esetére az alábbi eredményt kapjuk.

	v_a	v_b
u_a	5	0,89
u_b	0,89	2

Táblázat 2. Blb pontszámok hozzárendelések mérlegelése során.

Ez esetben látszik, hogy az aszimmetria lepontozása miatt a Blb algoritmus a megfelelő hozzárendeléseket fogja előnyben részesíteni. Mérési eredmények is igazolják [10], hogy a Blb algoritmus általában jelentősen jobb eredményeket tud elérni, mint a Nar algoritmus, vagy a többi korszerű algoritmus (a [22] cikk alapján).

Ezek az algoritmusok relatíve magas felidézéssel dolgoznak, de a hiba arányuk alacsony. A teljesség kedvéért bemutatjuk a **Korula-Lattazni algoritmust** (röviden: KL), amely magasabb felidézést képes bizonyos esetekben elérni, viszont a hiba aránya jóval magasabb [29].

A KL algoritmus is iteratív módon működik. Minden egyes iterációban kiválasztja a csúcsok egy részhalmazát, úgy, hogy iterációnként a legnagyobb fokszámúaktól kezdve bővíti a kiválasztott csúcsok körét. Párokat képez a két gráf valamennyi kiválasztott csúcsa között, és pontszámként hozzárendeli ezekhez a közös hozzárendelések számát. Ezek közül azokat választja ki új hozzárendelésre, ahol sem a háttérismeretből, sem az anonim gráfból származó csúcshoz sincs jobb hozzárendelés. Az így elérhető hozzárendeléseket már a következő iteráció során figyelembe veszi.

Az algoritmus ilyen módon valamennyi iteráció során felülvizsgálja a korábbi hozzárendeléseket, azokat javítva, amennyiben megfelelőbb hozzárendelési lehetőség adódik. Az iterációk száma a KL eljárásnál előre meghatározott érték, nem függ az aktuális iteráció sikerességétől.

4.4 Esettanulmány

Az alábbiakban mérésekre és de-anonimizációs támadás szimulációjára építve bemutatjuk az anonimizálási eljárások kiértékelésének módszertanját. Feltételezzük, hogy létezik egy olyan tranzakciós adatbázis, amit kapcsolati formában szeretne közzé tenni egy pénzügyi szereplő (megjegyezzük, hogy a kapcsolati forma könnyen táblázatos formátumúvá alakítható). Ezt vizsgáljuk meg az előző fejezetben ismertetett módszertan segítségével.

Az egyes alfejezetben – az áttekinthetőség kedvéért – jelöljük a módszertan vonatkozó lépéseinek számát is.

4.4.1 Védendő adathalmaz ismertetése (módszertan 1. lépése)

Mivel megfelelő struktúrájú és jellegű pénzügyi adatbázis nem áll rendelkezésre, ezért az Enron adathalmazt fogjuk a vizsgálathoz alkalmazni [30]. Ez egy levelezési adathalmaz, amely körülbelül félmillió emailből áll össze, és az amerikai Federal Energy Regulatory Commission tette közzé, amikor vizsgálatot folytatott le az Enron céggel szemben 2001-ben.

Az adathalmaz 36 692 címet (csúcsot) tartalmaz, amelyek között akkor szerepel él, ha a két cím között legalább egy emailt elküldtek valamelyik irányból. A gráfnak 183 831 éle van. Ezt az adathalmazt több kapcsolatrendszerre épülő de-anonimizálási tanulmány is alkalmazta, így a tématerülettől nem idegen a felhasználása [10][22][29].

4.4.2 Támadó modellezése (módszertan 2. lépése)

Az adatokat preventív módon tudjuk az anonimizálás segítségével védeni. Ebben nehezítés, hogy ehhez előre kellene ismernünk a rosszindulatú fél, a támadó szándékát, lehetőségeit és képességeit. Erre viszont nincs lehetőségünk. Ezért többféle képességű támadóra vizsgáljuk a támadások sikerességének lehetőségét, amely tulajdonságot az anonimizált adatra vonatkozó háttérismeret erősségével modellezünk.

Minél több erőforrás (például releváns adatok, számítási kapacitás és/vagy pénz) és szaktudás áll egy támadó rendelkezésére, annál pontosabb háttérismeretet tud építeni. Ez annál pontosabban fogja modellezni az anonimizált adatot és ezért annál nagyobb lesz az átfedése a két adathalmaznak.

Ezt úgy tudjuk modellezni, hogy a perturbációs eljárás segítségével többféle lehetséges háttérismeret és anonimizált adatpárt hozunk létre. Ehhez a NarPert eljárást alkalmazzuk, és a következő eltérő erősségű támadókat modellezzük:

- *Gyenge támadó (GyT).* A létrehozott adatoknál $\alpha_v = 0{,}25$ és $\alpha_e = 0{,}5$ értékeket alkalmazunk. A létrehozott gráfok 14 ezer csúcsot és 45-50 ezer élt tartalmaznak. A háttérismeret és az anonimizálandó adat között az átfedés 4 306 csúcs.
- *Közepesen erős támadó (KT).* A létrehozott adatoknál $\alpha_v = 0{,}5$ és $\alpha_e = 0{,}75$ értékeket alkalmazunk. A létrehozott gráfok kb. 20 ezer csúcsot és kb. 85 ezer élt tartalmaznak. A háttérismeret és az anonimizálandó adat között az átfedés 12 119 csúcs.
- *Erős támadó (ET).* A létrehozott adatoknál $\alpha_v = 0{,}75$ és $\alpha_e = 0{,}9$ értékeket alkalmazunk. A létrehozott gráfok kb. 27 ezer csúcsot és kb. 131 ezer élt tartalmaznak. A háttérismeret és az anonimizálandó adat között az átfedés 21 507 csúcs.

Ezzel a három támadóval vizsgáljuk tovább az anonimizálási eljárások erősségét.

4.4.3 Anonimizációs eljárások és hasznosság (módszertan 3. lépése)

A korábban tárgyalt anonimizálási eljárásokat alkalmaztuk, a szakirodalomban található eredmények alapján meghatározva a szükséges paramétereket [10][22]. A *Switch(k)* eljáráshoz a *k*=10 paramétert választottuk, ami a gráfban lévő élek 10%-ának áthelyezését jelenti. A *k-DA* eljáráshoz a k=50 beállítást használtuk, amely azt jelenti, hogy a csúcsok fokszáma alapján, élek hozzáadása által éri el a k-anonimitást k=50-es értékre. Végezetül pedig a DP(ε) jelölésű, differenciális adatvédelemre épülő anonimizálási eljárást alkalmaztuk $\varepsilon = 50$ paraméterrel, ami – az előző két esethez hasonlóan – egy erősebb anonimizálási szintnek felel meg a szakirodalmi eredmények iránymutatása alapján.

Az anonimizálás eredményességét meghatározó két fő mutató az adatok hasznosítása, a másik az anonimizálás újra-azonosítással szembeni ellenálló képessége. A hasznosságot az adott alkalmazás határozza meg, amelyet a jelen tanulmányban nincs lehetőségünk konkretizálni. Ezért az alább látható táblázatban foglaljuk össze néhány általános metrika, mint hasznosítási szempont alapján az anonimizálás adat módosítási hatását, ami közvetlen hatással lehet a hasznosításra. Az újra-azonosítással szembeni ellenállóságot pedig a következő fejezetben vizsgáljuk.

	Switch(k) *(k=10)*			*k-DA* *(k=50)*			DP(ε) ($\varepsilon = 50$)		
Támadó erősség	GyT	KT	ET	GyT	KT	ET	GyT	KT	ET
Fokszámeloszlás korrelációja	0,999	0,999	0,999	0,999	0,999	0,999	0,975	0,993	0,984
LCC	0,915	0,899	0,884	0,961	0,966	0,966	0,816	0,950	0,868
BC	0,977	0,969	0,975	0,867	0,882	0,885	0,889	0,828	0,872

Táblázat 3. Strukturális módosulás mértéke anonimizálás előtti és utáni állapotok között. (LCC: Local Clustering Coefficient; BC: Betweenness Centrality)

Az eredmények alapján az anonimizálási eljárások alkalmazása ellenére az adatok valószínűleg jól használhatóak maradtak.

4.4.4 De-anonimizáció beállításai és futtatása (módszertan 4-5. lépései)

A korábban bemutatott, korszerű de-anonimizációs algoritmusokkal ellenőriztük az anonimizáció erősségét. Az algoritmusok paramétereit a szakirodalom eredményei alapján határoztuk meg [10]. A cél, hogy az algoritmusok kellően mohó viselkedést tanúsítsanak, de lehetőleg elfogadhatóan alacsony hiba mellett.

A Nar algoritmust $\theta = 0{,}1$ paraméter beállítás mellett alkalmaztuk, ami a szakirodalomban szereplő mérésekben ökölszabályként 4-5% körüli hibát adott. A Blb algoritmust $\theta = 0{,}1$ és $\delta = 0{,}5$ paraméterekkel alkalmaztuk, amelyek hasonlóan a Nar paraméterezéséhez, jellemzően magas felidézés és alacsony hibaértékeket adtak a szakirodalomban publikált mérésekben. A KL algoritmusnál a SecGraph könyvtár implementációját használtuk [22], amely egy paramétert enged, hogy az adott iterációból hány párosítást tarthat meg az algoritmus. Ezt méréseink alapján 2000-re állítottuk.

Valamennyi algoritmus inicializálásához szükséges megadni egy kezdeti hozzárendelést (μ_0). Általános szabály, hogy bár többféleképpen kiválaszthatjuk szimulációs vizsgálatokhoz, de jellemzően a magasabb fokszámú csúcsokkal hatékonyabban működik az algoritmus. A fokszám melletti további szempontok is jelentősen befolyásolják, hogy összesen hány hozzárendelésre van szükségünk ahhoz, hogy az algoritmus el tudjon indulni [28].

A leghatékonyabb például, ha a legmagasabb fokszámú csúcsokat választjuk, de szintén elég hatékony, ha a gráf fokszám szerinti felső 1%-ából választunk. Legfeljebb a terjedési fázis elindulásához több hozzárendelés kell majd, például [10]-es cikk 5. ábra alapján Blb-hez $|\mu_0| = 1$ helyett 3-4, Nar esetén pedig kb. $|\mu_0| = 60$ helyett 70-nél is több.

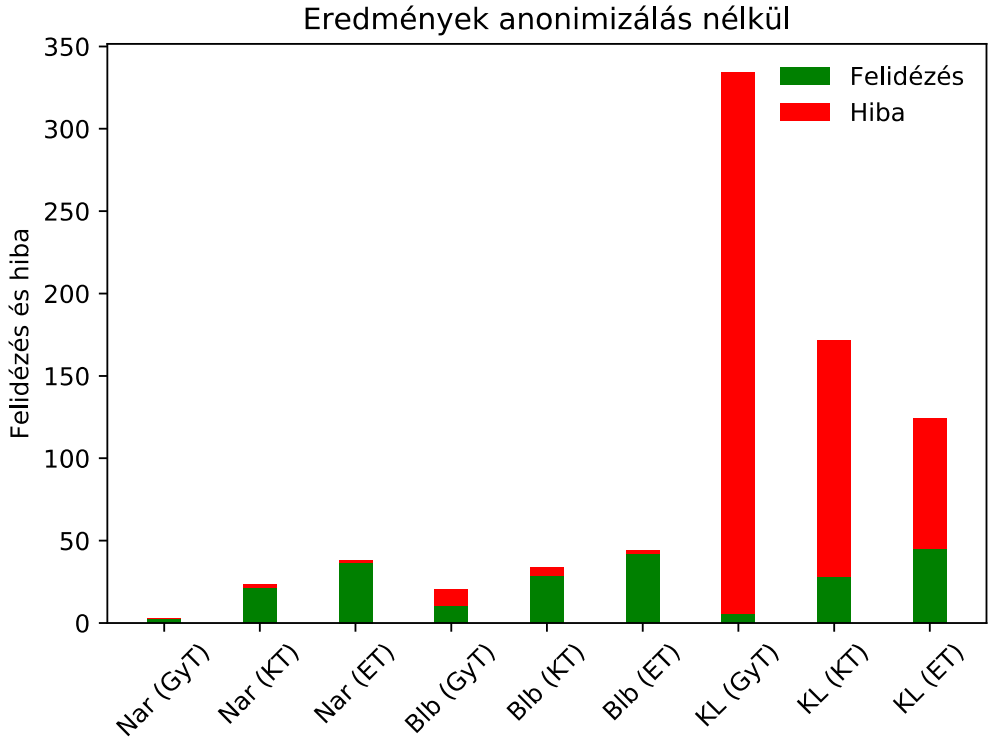
Jelen tanulmányunkban, az alapigazságban szereplő, 100 legmagasabb fokszámú csúcsot használjuk a kezdeti hozzárendelés kialakítására. Részünkről ez egy elfogadható egyszerűsítés a tanulmány áttekinthetősége és érthetősége kedvéért, de egy vizsgálat során ez is a vizsgálat tárgya kellhet, hogy legyen: egy ésszerűen feltételezhető támadó milyen inicializálási eljárásokkal rendelkezhet?

Ezekkel a paraméterekkel lefuttattuk az algoritmusokat az anonim adathalmazokra, illetve a csupán perturbációval rendelkezőekre. A KL futtatásához a

SegGraph keretrendszert használtuk [22], míg a Nar és Blb futtatásához az SALab-ot [31].

4.4.5 Kiértékelés (módszertan 6. lépése)

Először a nem anonimizált, perturbációs eljárással készített adatokon futtattuk le az újra-azonosítási algoritmusokat. Az eredményeket bemutatjuk a 11. ábrán. Ezek a mérési eredmények két fontos következtetést engednek meg a további vizsgálatok szempontjából. Egyrészt, hogy bár a KL algoritmus a felidézés szempontjából valóban képes tartani az élmezőnnyel a versenyt, ezekhez a felidézés értékekhez olyan magas számú hibás hozzárendelések társulnak, hogy a végeredmény használhatatlan lesz. Másrészt, hogy nagyságrendileg hasonló hiba arány mellett a Blb jelentősen magasabb felidézést képest elérni, mint a Nar. A KL-t mellőzzük a továbbiakban, és a Blb, Nar algoritmusokkal dolgozunk tovább.



Ábra 11. Anonimizáció nélküli, perturbációval előállított adatokon elért de-anonimizálási eredmények.

Az anonimizálásra vonatkozó eredményeket a 4. táblázat összesíti. Ezek alapján megállapíthatjuk, hogy gyenge támadó (GyT) esetén az anonimizálási eljárások megfelelőek, de ennél erősebb esetben nem, ugyanis a Blb algoritmus az anonimitás

kompromittálásához, és így személyes adat szivárgásához vezethet. Például a legerősebb védekezési eljárás közepes támadó esetén a differenciális adatvédelem, de még ebben az esetben is a felhasználók 14,11%-ának az adatai kiszivárognak.

Ez azért relevánsabb eredmény, mint az erős támadó eredménye, mert a valóságban egy közepes támadó kockázata magasabb, mert nagyobb valószínűséggel fordulhat elő. Amennyiben az adott kontextusban nem zárható ki kellően hihetően egy közepes vagy erős támadó lehetősége, megállapíthatjuk, hogy egy további, erősebb megoldásra van szükség, ami ezekben az esetekben is kizárja a de-anonimizálás lehetőségét.

		Nar		**Blb**	
		Felidézés (%)	**Hiba (%)**	**Felidézés (%)**	**Hiba (%)**
GyT	***Switch(k) (k=10)***	2,67	0,25	6,99	11,91
	k-DA (k=50)	2,69	0,25	7,64	8,15
	DP(ε) ($\varepsilon = 50$)	2,41	0,11	3,08	3,57
KT	***Switch(k) (k=10)***	1,59	0,25	23,15	6,41
	k-DA (k=50)	10,16	1,15	24,63	5,84
	DP(ε) ($\varepsilon = 50$)	1,24	0,15	14,11	7,73
ET	***Switch(k) (k=10)***	25,88	1,86	35,07	3,99
	k-DA (k=50)	30,65	1,39	39,04	2,95
	DP(ε) ($\varepsilon = 50$)	22,28	0,74	26,67	6,23

Táblázat 4. De-anonimizálási eredmények különféle erősségű támadók és anonimizálási eljárások esetén.

Megjegyezzük, hogy lehetnek olyan helyzetek, amikor a közepes és erős támadók ugyan valószínűsíthetők, de a 20-25%-os újra-azonosítási arány esetén az információ szivárgás hatása nem jelent komoly kockázatot az adathalmaz egészére nézve, míg, ha ez

50-60% lenne, az jelentősebb veszélyt jelentene (ez az adat típusától, lehetséges hatásától függ).

Mivel a DP(ε) eljárás bizonyult a legerősebbnek, célszerű ezzel tovább kísérletezni. A szakirodalom által javasolt $\varepsilon = 50$ paramétert próbáljuk meg először $\varepsilon = 25$-re állítani. Az alábbi táblázat foglalhatjuk össze a hasznosságát ennek az anonim adatnak a korábbi metrikák alapján.

	GyT	**KT**	**ET**
Fokszámeloszlás korrelációja	0,871	0,976	0,958
LCC	0,731	0,732	0,726
BC	0,913	0,880	0,907

Táblázat 5. Differenciális adatvédelem hasznossági metrikái különféle támadókkal szemben $\varepsilon = 25$ paraméterre.

Ezek az eredményiek azt mutatják, hogy az adat hasznossága megfelelő lehet, tehát ebből a szempontból az anonimizálási eljárást elfogadhatjuk. A de-anonimizálási támadások eredményeit a 6. táblázat foglalja össze.

	Nar		**Blb**	
	Felidézés (%)	**Hiba (%)**	**Felidézés (%)**	**Hiba (%)**
GyT	2,46	0,16	2,76	2,58
KT	0,93	0,1	1,06	0,99
ET	0,82	0,05	0,83	0,62

Táblázat 6. De-anonimizálási eredmények $\varepsilon = 25$ paraméterű differenciális adatvédelemmel.

Látható, hogy a támadások a kezdeti hozzárendeléshez kevés új hozzárendelést tudtak hozzáadni. Erre a beállításra megállapítható, hogy az anonimizálási eljárás megfelelő: az adat hasznossága elégséges szintű maradt, de az anonimizálás korszerű de-anonimizálási eljárásokkal nem kompromittálható.

5 Összefoglalás

Jelen tanulmány fő kérdésköre a tranzakciós adatok anonimitása. A tanulmány első részében bemutattuk a személyes (azonosított) és az anonim adatok közötti ellentétet, és az anonimizálás alapjait tárgyaltuk klasszikus, illetve nagy dimenzionalitású esetben. Erre az alapra építve bevezettük a nagy méretű adatbázisok de-anonimizálási elveit.

A tanulmány egyik kulcs gondolata, hogy a tranzakciós adatok anonimitási vizsgálatát úgy érdemes végrehajtani, hogy az ilyen adatokat gráfokká konvertáljuk és a vonatkozó de-anonimizációs szakirodalom eljárásait alkalmazzuk. Ehhez kidolgoztunk egy módszertant, ami a GDPR szemüvegén keresztül nyújt segítséget ahhoz, hogy egy bizonyos anonimizálási eljárás hatékonyságát ellenőrizhessük.

A tanulmány végén megadtuk azokat a technikai eszközöket, amelyek a módszertan végrehajtásához szükségesek: bemutattunk korszerű gráf anonimizálási és újra-azonosítási eljárásokat. Mindezek működését egy esettanulmányon demonstráltuk, ahol egy tranzakciós adatkezelő szerepében azt az anonimizálási eljárást kerestük meg, amelynek a segítségével a tranzakciós adatok használhatók maradnak, de az adatokban lévő rekordok nem lesznek újra azonosíthatók.

Köszönetnyilvánítás

A könyv/cikk/tanulmány a Magyar Nemzeti Bank és a Budapesti Műszaki és Gazdaságtudományi Egyetem között létrejött Együttműködés keretében és finanszírozásával készült a Digitalizáció, mesterséges intelligencia és adatkorszak műhelyben.

Irodalomjegyzék

[1] GDPR. 2018 reform of EU data protection rules. European Comission. May 25 2018.

[2] A. Narayanan and V. Shmatikov, "De-anonymizing Social Networks," 2009 30th IEEE Symposium on Security and Privacy, 2009, pp. 173-187, doi: 10.1109/SP.2009.22.

[3] Sweeney L. (1998) Datafly: a system for providing anonymity in medical data. In: Lin T.Y., Qian S. (eds) Database Security XI. IFIP Advances in Information and Communication Technology. Springer, Boston, MA

[4] Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco), 671*, 1-34.

[5] Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. 2015092903. September 28, 2015.

[6] Latanya Sweeney, „k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10. (2002) Issue 5., pp. 557-570.

[7] Latanya Sweeney, "k-anonymity: a model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10. (2002) Issue 5., pp. 557-570.

[8] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, 2008, pp. 111-125, doi: 10.1109/SP.2008.33.

[9] Rocher, L., Hendrickx, J.M. & de Montjoye, YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10, 3069 (2019).

[10] Gulyás, G. G., Simon, B., & Imre, S. (2016, October). An efficient and robust social network de-anonymization attack. In Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society (pp. 1-11).

[11] Huy Pham, Cyrus Shahabi, and Yan Liu. 2013. EBM: an entropy-based model to infer social strength from spatiotemporal data. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13). Association for Computing Machinery, New York, NY, USA, 265–276.

[12] Hern, A. ‘Anonymous’ browsing data can be easily exposed, researchers reveal. The Guardian (1 Aug 2017).

[13] Fox-Brewster, T. 120 million american households exposed in ‘massive’ConsumerView database leak. Forbes (2017).

[14] C. C. Aggarwal, A. Hinneburg, D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. Proceedings of the ICDT Conference, pp. 420-434, 2001.

[15] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques (2014)

[16] Arvind Narayanan: Eccentricity Explained. https://33bits.wordpress.com/2008/10/03/eccentricity-explained/

[17] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (March 2007). L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1 (2007), 3. DOI:https://doi.org/10.1145/1217299.1217302

[18] Li, N., Li, T., and Venkatasubramanian, S. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, 2007, pp. 106-115.

[19] McSherry F. and Talwar, K. "Mechanism Design via Differential Privacy," 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), Providence, RI, 2007, pp. 94-103

[20] Public Use Microdata Sample (PUMS) 5-percent https://www.census.gov/main/www/pums.htm

[21] Rocher, L., Hendrickx, J.M. & de Montjoye, YA. „Too unique to hide" https://cpg.doc.ic.ac.uk/individual-risk/

[22] Shouling Ji, Weiqing Li, Prateek Mittal, Xin Hu, Raheem Beyah: SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization. Usenix Security 2015.

[23] Yartseva, Lyudmila, Matthias Grossglauser: On the performance of percolation graph matching. Proceedings of the first ACM conference on Online social networks. ACM, 2013.

[24] Pedram Pedarsani, Daniel R. Figueiredo, Matthias Grossglauser: A Bayesian method for matching two similar graphs without seeds. 51st Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, 2013.

[25] Xiaowei Ying and Xintao Wu: Randomizing Social Networks: a Spectrum Preserving Approach. Proceedings of the 2008 SIAM International Conference on Data Mining. 2008.

[26] Kun Liu and Evimaria Terzi. 2008. Towards identity anonymization on graphs. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008.

[27] Alessandra Sala, Xiaohan Zhao, Christo Wilson, Haitao Zheng, and Ben Y. Zhao. 2011. Sharing graphs using differentially private graph models. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement, 2011.

[28] G. G. Gulyás and S. Imre, "Measuring importance of seeding for structural de-anonymization attacks in social networks," 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), 2014, pp. 610-615, doi: 10.1109/PerComW.2014.6815276.

[29] Korula, Nitish, and Silvio Lattanzi. "An efficient reconciliation algorithm for social networks." arXiv preprint arXiv:1307.1690 (2013).

[30] Klimt, Bryan, and Yiming Yang. "Introducing the Enron corpus." In CEAS. 2004.

[31] G. G. Gulyás: „SALab: Framework for the analysis of structural de-anonymization algorithms.” (2016), GitHub repository, https://github.com/gaborgulyas/salab/

Csarnó Tamás Péter hallgatóként végzi M.Sc-s tanulmányait a Budapesti Műszaki és Gazdaságtudományi Egyetem, Villamosmérnöki és Informatikai kar mechatronikai mérnöki mesterképzési szakán. Korábbi tanulmányai során kutatómunkát folytatott adatvédelem és gépi tanulás témakörökben, azon belül arcfelismerő modellekkel kapcsolatos adatvédelmi kérdésekben. Jelenleg diplomamunkáját írja gépi tanulási eljárásokkal generált adatok adatvédelmi vizsgálata címen.

Gulyás Gábor György 2015-ben szerzett PhD fokozatott a Budapesti Műszaki és Gazdaságtudományi Egyetem Hálózati Rendszerek és Szolgáltatások Tanszékén. 2015 és 2018 között poszdoktori kutató és kutatómérnök volt a Privatics csapatban az INRIA-nál (Franciaország); 2019 óta tudományos munkatárs a BME Automatizálási és Alkalmazott Informatikai Tanszéken. Az adatvédelem informatikai kérdéseivel 2005 óta foglalkozik aktívan. Főbb szakmai érdeklődési és kutatási területei: anonimizálás és de-anonimizálás, webes megfigyelés adatvédelmi kérdései, gépi tanulás adatvédelmi kihívásai és a GDPR szabályozás.