PRACOVNÍ SKUPINA PRO OCHRANU ÚDAJŮ ZŘÍZENÁ PODLE ČLÁNKU 29

0829/14/CS
WP216

Stanovisko č. 5/2014 k technikám anonymizace

Přijaté dne 10. dubna 2014

PRACOVNÍ SKUPINA PRO OCHRANU FYZICKÝCH OSOB V SOUVISLOSTI SE ZPRACOVÁNÍM OSOBNÍCH ÚDAJŮ

zřízená směrnicí Evropského parlamentu a Rady 95/46/ES ze dne 24. října 1995,

s ohledem na články 29 a 30 této směrnice,

s ohledem na svůj jednací řád,

PŘIJALA TOTO STANOVISKO:

SHRNUTÍ

Pracovní skupina analyzuje v tomto stanovisku účinnost a hranice stávajících technik anonymizace na základě právních předpisů EU o ochraně údajů a uvádí doporučení, jak s těmito technikami nakládat, přičemž bere v potaz zbytkové riziko identifikace, jemuž se u žádné z těchto technik nelze vyhnout.

Pracovní skupina si uvědomuje potenciální hodnotu anonymizace, zejména coby strategie, jak obecně využívat výhod „přístupných údajů“ pro jednotlivce i společnost jako celek a zároveň zmírnit rizika pro dotčené fyzické osoby. Případové studie a výzkumné publikace nicméně ukázaly, jak je obtížné vytvořit skutečně anonymní soubor údajů a zároveň zachovat tolik základních informací, kolik je jich k danému úkolu potřeba.

Podle směrnice 95/46/ES a dalších příslušných právních nástrojů EU je anonymizace výsledkem zpracování osobních údajů tak, aby se nezvratně zabránilo identifikaci. Správci by současně měli přihlédnout k několika skutečnostem s ohledem na všechny prostředky, o nichž lze „důvodně předpokládat“, že mohou být použity pro identifikaci (jak správcem, tak kteroukoli třetí stranou).

Anonymizace představuje další zpracování osobních údajů a jako taková musí splňovat požadavek na slučitelnost s přihlédnutím k právním důvodům a okolnostem dalšího zpracování. Anonymizované údaje kromě toho sice nespadají do oblasti působnosti právních předpisů na ochranu údajů, ale subjekty údajů mohou mít nadále nárok na ochranu podle jiných předpisů (například předpisů na ochranu důvěrné povahy komunikací).

V tomto stanovisku jsou popsány hlavní techniky anonymizace, zejména randomizace a generalizace. Toto stanovisko pojednává konkrétně o metodách noise addition, permutace, diferenciální utajení, agregace, k-anonymita, l-diverzita a t-blízkost. Vysvětluje jejich princip, jejich silné a slabé stránky i obvyklé chyby a nedostatky spojené s používáním každé z technik.

Ve stanovisku je rozebrána spolehlivost každé techniky na základě tří kritérií:

(i) Je i nadále možné vyčlenit jednotlivou osobu?

(ii) Je nadále možné uvést do vzájemné souvislosti záznamy týkající se jedné fyzické osoby?

(iii) Je možné vyvodit informace týkající se jedné fyzické osoby?

Znalost hlavních silných a slabých stránek každé techniky napomáhá k rozhodnutí o tom, jak v daných souvislostech naplánovat vhodný proces anonymizace.

Za účelem objasnění některých skrytých nebezpečí a mylných představ se stanovisko věnuje i pseudonymizaci. Pseudonymizace není metodou anonymizace. Pseudonymizace pouze omezuje možnosti uvedení souboru údajů do souvislosti s původní identitou subjektu údajů a představuje tak užitečné bezpečnostní opatření.

Stanovisko dochází k závěru, že techniky anonymizace mohou poskytnout záruky utajení a mohou být použity tak, aby vznikly účinné procesy anonymizace, avšak pouze v případě, že je jejich použití náležitě promyšleno, což znamená, že je nutné jasně stanovit nutné předpoklady (souvislosti) a cíl (cíle), aby mohlo být dosaženo cílené anonymizace a současně byly

vyprodukovány užitečné údaje. O optimálním řešení by se mělo rozhodovat případ od případu, eventuálně za současného použití různých technik a s přihlédnutím k praktickým doporučením uvedeným v tomto stanovisku.

Správci údajů by nakonec měli vzít v úvahu, že anonymizovaný soubor údajů může i tak představovat pro subjekty údajů zbytkové riziko. Anonymizace a zpětná identifikace jsou totiž na jedné straně aktivními oblastmi výzkumu a pravidelně se zveřejňují nová zjištění. Na straně druhé mohou být i anonymizované údaje, např. statistiky, použity k obohacení stávajících profilů fyzických osob, čímž v oblasti ochrany údajů vyvstávají nové problémy. Na anonymizaci tudíž nelze pohlížet jako na jednorázový úkon a správci údajů by měli související rizika pravidelně přehodnocovat.

1 Úvod

Zařízení, čidla a sítě produkují velké objemy a nové druhy údajů a náklady na uchovávání údajů se stávají zanedbatelnými a současně vzrůstá zájem veřejnosti o tyto údaje a poptávka po jejich opětovném využití. „Přístupné údaje“ mohou pro společnost, jednotlivce a organizace představovat jednoznačný přínos, ale pouze tehdy, jsou-li respektována práva každé osoby na ochranu jejích osobních údajů a soukromého života.

Anonymizace může být dobrou strategií, jak zachovat výhody a snížit rizika. Jakmile je soubor údajů skutečně anonymizován a jednotlivé osoby již nelze identifikovat, evropský právní předpis o ochraně údajů se nepoužije. Z případových studií a výzkumných publikací je však zřejmé, že vytvoření skutečně anonymního souboru údajů z bohatého souboru osobních údajů při zachování co největšího množství základních informací tak, jak vyžaduje daný úkol, není snadnou záležitostí. Soubor údajů považovaný za anonymní může být například zkombinován s jiným souborem údajů tak, že lze identifikovat jednu nebo více fyzických osob.

Pracovní skupina analyzuje v tomto stanovisku účinnost a hranice stávajících technik anonymizace na základě právních předpisů EU o ochraně údajů a uvádí doporučení, jak tyto techniky pro vytvoření procesu anonymizace obezřetně a zodpovědně využívat.

2 Definice a právní analýza

2.1 Definice v právním kontextu EU

Směrnice 95/46/ES odkazuje na anonymizaci v 26. bodě odůvodnění, v němž anonymizované údaje vylučuje z oblasti působnosti právních předpisů o ochraně údajů:

> *„vzhledem k tomu, že zásady ochrany se musí vztahovat na veškeré informace týkající se identifikované či identifikovatelné osoby; že pro určení, zda je osoba identifikovatelná, je třeba přihlédnout ke všem prostředkům, které mohou být rozumně použity jak správcem, tak jakoukoli jinou osobou pro identifikaci dané osoby; že zásady ochrany se nevztahují na údaje, které byly anonymizovány tak, že subjekt údajů již není identifikovatelný; že kodexy chování ve smyslu článku 27 mohou být užitečným nástrojem pro poskytování informací o způsobech, kterými mohou být údaje anonymizovány a uchovány v podobě, která již neumožňuje identifikaci subjektu údajů;“*[1]

Podrobné čtení 26. bodu odůvodnění přináší definici pojmu anonymizace. Ustanovení 26. bodu odůvodnění uvádí, že pro anonymizaci jakýchkoli údajů musí být údaje zbaveny dostatečného množství prvků tak, aby subjekt údajů již nemohl být identifikován. Přesněji řečeno, údaje musí být zpracovány tak, aby již nemohly být použity k identifikaci fyzické osoby, a to za využití všech prostředků, které mohou být rozumně použity jak správcem, tak třetí stranou. Důležitým faktorem je skutečnost, že zpracování musí být nevratné. Tato

[1] Kromě toho je třeba uvést, že tento přístup se uplatňuje i v návrhu nařízení EU o ochraně údajů, kde se v 23. bodě odůvodnění uvádí, že: „při určování, zda je osoba identifikovatelná, by se mělo přihlédnout ke všem prostředkům, o nichž lze důvodně předpokládat, že je správce nebo jakákoli jiná osoba použijí pro identifikaci dané osoby“.

směrnice nevysvětluje, jak by měl nebo může být tento proces zabránění identifikaci proveden[2]. Soustředí se na výsledek: údaje musí mít takovou podobu, aby znemožňovaly identifikovat subjekt údajů jakýmikoli prostředky, „které mohou být rozumně použity". Odvolává se na kodexy chování jakožto nástroj ke stanovení mechanismů anonymizace a uchovávání v podobě, která „již neumožňuje" identifikaci subjektu údajů. Směrnice tak ě zavádí velmi vysoký standard.

Směrnice o ochraně soukromí v odvětví elektronických komunikací (směrnice 2002/58/ES) rovněž ve stejném smyslu odkazuje na „anonymizaci" a „anonymní údaje". Ustanovení 26. bodu odůvodnění uvádí:

> *„Provozní údaje používané pro marketing komunikačních služeb nebo pro poskytování služeb s přidanou hodnotou by měly být po poskytnutí služby vymazány nebo anonymizovány."*

V souladu s tím čl. 6 odst. 1 stanoví, že:

> *„Provozní údaje vztahující se k účastníkům a uživatelům zpracovávané a uchovávané provozovatelem veřejné komunikační sítě nebo poskytovatelem veřejně dostupné služby elektronických komunikací, musí být vymazány nebo anonymizovány, jakmile již nejsou potřebné pro přenos sdělení, aniž by tímto byly dotčeny odstavce 2, 3 a 5 tohoto článku a čl. 15 odst. 1."*

Dále čl. 9 odst. 1 stanoví:

> *„Mohou-li být zpracovávány lokalizační údaje odlišné od provozních údajů, které se vztahují k uživatelům nebo účastníkům veřejných komunikačních sítí nebo veřejně dostupných služeb elektronických komunikací, je možné tyto údaje zpracovávat pouze poté, co byly anonymizovány údaje, anebo se souhlasem uživatelů nebo účastníků v nezbytném rozsahu a po nezbytnou dobu pro poskytování služeb s přidanou hodnotou."*

To vychází z logické úvahy, že výsledek anonymizace jakožto použité techniky by za stávajícího stavu technologií měl být stejně trvalý jako smazání, tzn. znemožňovat zpracování osobních údajů[3].

2.2 Právní analýza

Analýza formulací týkající se anonymizace v hlavních nástrojích EU na ochranu údajů umožňuje vyzdvihnout čtyři hlavní znaky:

- Anonymizace může být výsledkem zpracování osobních údajů s cílem nevratně zabránit identifikaci subjektu údajů.

[2] Tento pojem je dále rozpracován na straně 8 tohoto stanoviska.

[3] Je třeba připomenout, že anonymizace je definována i v mezinárodních normách, např. v normě ISO 29100, a to jako „proces, jímž se informace, podle nichž je možno identifikovat osobu (PII), nevratně mění tak, že hlavní část těchto informací již nemůže přímo ani nepřímo identifikovat ani správce osobních informací sám, ani ve spolupráci s jiným subjektem" (ISO 29100:2011). Nevratnost změny provedené v osobních údajích tak, aby byla znemožněna přímá nebo nepřímá identifikace, je klíčová i pro ISO. Z tohoto hlediska se norma značně blíží zásadám a myšlenkám, jež jsou základem směrnice 95/46/ES. To se týká i definic, jež lze nalézt v některých vnitrostátních právních předpisech (například v Itálii, v Německu a ve Slovinsku), které kladou důraz na neidentifikovatelnost a poukazují na „nepřiměřené úsilí", jež by bylo třeba vynaložit ke zpětné identifikaci (Německo, Slovinsko). Francouzský zákon na ochranu osobních nicméně stanoví, že údaje zůstávají osobními údaji, i když je mimořádně obtížné a nepravděpodobné, aby byl objekt údajů zpětně identifikován. Neexistuje zde tedy ustanovení o otázce „přiměřenosti".

- Lze předpokládat použití několika technik anonymizace, v právních předpisech EU není předepsána žádná norma.

- Význam je třeba přikládat prvkům vyplývajícím ze souvislostí: je nutno přihlédnout ke „všem“ prostředkům, které mohou být k identifikaci „rozumně použity“ správcem nebo třetími stranami, přičemž je třeba věnovat zvláštní pozornost tomu, co může být na základě stávajícího stavu technologií „rozumně použito“ (vzhledem k nárůstu možností výpočetní techniky a dostupných nástrojů).

- Nedílnou součástí anonymizace je rizikový faktor: tento faktor je třeba zvažovat při posuzování opodstatněnosti jakékoli techniky anonymizace, včetně možného použití jakýchkoli údajů, které jsou touto technikou „anonymizovány“, přičemž by měla být posouzena závažnost a pravděpodobnost tohoto rizika.

V tomto stanovisku se místo pojmů „anonymita“ nebo „anonymní údaje“ používá pojem „technika anonymizace“, aby se zdůraznilo nedílné zbytkové riziko zpětné identifikace, které je spojeno s každým technicko-organizačním opatřením, jehož cílem je „anonymizace“ údajů.

2.2.1 Zákonnost procesu anonymizace

Zaprvé, anonymizace je technika, která se používá na osobní údaje tak, aby bylo dosaženo nevratného znemožnění identifikace. Prvotním předpokladem je tudíž shromáždění a zpracování osobních údajů v souladu s platnými právními předpisy o uchovávání údajů v identifikovatelném formátu.

V této souvislosti je proces anonymizace, tedy zpracovávání osobních údajů za účelem dosažení anonymizace, případem „dalšího zpracování“. Jako takové musí toto zpracování vyhovět testu slučitelnosti podle pokynů stanovených pracovní skupinou v jejím stanovisku č. 3/2013 k účelovému omezení[4].

To znamená, že právní základ pro anonymizaci lze v zásadě nalézt v jakémkoli důvodu uvedeném v článku 7 (včetně oprávněného zájmu správce) za předpokladu, že jsou splněny i požadavky na kvalitu údajů uvedené v článku 6 směrnice, a s přihlédnutím ke zvláštním okolnostem a všem faktorům uvedeným ve stanovisku pracovní skupiny k účelovému omezení[5].

Na druhou stranu je třeba poukázat na ustanovení čl. 6 odst. 1 směrnice 95/46/ES (ale i čl. 6 odst. 1 a čl. 9 odst. 1 směrnice o ochraně soukromí v odvětví elektronických komunikací), neboť dokládají potřebu uchovávat osobní údaje „ve formě umožňující identifikaci“ po dobu ne delší, než je nezbytné pro uskutečnění cílů, pro které jsou shromažďovány nebo dále zpracovávány.

---

[4] Stanovisko pracovní skupiny zřízené podle článku 29 č. 3/2013, dostupné na: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5] To zejména znamená, že musí být provedeno věcné posouzení s přihlédnutím ke všem důležitým okolnostem a se zvláštním zřetelem k následujícím klíčovým faktorům:
a) vztah mezi účelem, pro nějž byly shromážděny osobní údaje, a účelem dalšího zpracování;
b) souvislosti, v nichž byly osobní údaje shromažďovány, a přiměřená očekávání subjektů údajů, pokud jde o jejich další použití;
c) povaha osobních údajů a dopad dalšího zpracovávání na subjekty údajů;
d) záruky přijaté správcem za účelem zajištění spravedlivého zpracování a zabránění přílišnému dopadu na subjekty údajů.

Toto ustanovení je samo o sobě jasným argumentem pro to, aby byly osobní údaje anonymizovány přinejmenším jako výchozí možnost (na základě různých právních požadavků, například těch, jež jsou uvedeny ve směrnici o ochraně soukromí v odvětví elektronických komunikací v souvislosti s provozními údaji). Pokud chce správce uchovávat tyto osobní údaje i poté, co bylo dosaženo cíle jejich původního nebo dalšího zpracování, měly by být použity techniky anonymizace, aby se nevratně zabránilo identifikaci.

Pracovní skupina se tedy domnívá, že anonymizaci jakožto případ dalšího zpracování osobních údajů lze považovat za slučitelnou s původními cíli zpracování, ale pouze za podmínky, že procesem anonymizace spolehlivě vzniknou anonymizované informace ve smyslu uvedeném v tomto dokumentu.

Je třeba rovněž zdůraznit, že anonymizace musí být provedena v souladu s právními omezeními, na něž poukázal Evropský soudní dvůr ve věci C-553/07 (*College van burgemeester en wethouders van Rotterdam v. M.E.E. Rijkeboer*), jež se týká potřeby uchovávat údaje v identifikovatelném formátu, například proto, aby subjekty údajů mohly vykonávat své právo na přístup. ESD rozhodl, že: „*Článek 12 písm. a) směrnice [95/46] ukládá členským státům, aby stanovily právo na přístup k informacím o příjemcích nebo kategoriích příjemců údajů, jakož i k obsahu sdělovaných informací platící nejen ohledně přítomnosti, ale i minulosti. Členským státům přísluší stanovit lhůtu pro uchovávání těchto informací, jakož i odpovídající přístup k nim, které zajistí spravedlivou rovnováhu mezi jednak zájmem subjektu údajů na ochraně jeho soukromí, zejména uplatněním práv a opravných prostředků stanovených směrnicí, a jednak zátěží, kterou pro správce představuje povinnost tyto informace uchovávat.*“

To je zvláště důležité vzhledem k tomu, že pokud jde o anonymizaci, správce vychází z čl. 7 písm. f) směrnice 95/46: oprávněný zájem správce údajů musí být vždy v rovnováze s právy a se základními svobodami subjektů údajů.

Například vyšetřování nizozemského orgánu pro ochranu údajů v letech 2012–2013, které se týkalo technologií hloubkové kontroly paketů používaných čtyřmi mobilními operátory, ukázalo, že podle čl. 7 písm. f) směrnice 95/46 existuje právní důvod pro anonymizaci obsahu provozních údajů co nejdříve po jejich shromáždění. Článek 6 směrnice o ochraně soukromí v odvětví elektronických komunikací stanoví, že provozní údaje vztahující se k účastníkům a uživatelům, jež jsou zpracovávány a uchovávány provozovatelem veřejné komunikační sítě nebo poskytovatelem veřejně dostupné služby elektronických komunikací, musí být vymazány nebo anonymizovány co nejdříve. V tomto případě, jelikož je to umožněno článkem 6 směrnice o ochraně soukromí v odvětví elektronických komunikací, existuje odpovídající právní důvod v článku 7 směrnice o ochraně údajů. Lze to vyjádřit i obráceně: pokud není způsob zpracování povolen článkem 6 směrnice o ochraně soukromí v odvětví elektronických komunikací, nemůže v článku 7 směrnice o ochraně údajů existovat právní důvod.

2.2.2 Možná identifikovatelnost anonymizovaných údajů

Pracovní skupina se pojetí osobních údajů podrobně věnovala ve stanovisku č. 4/2007 k osobním údajům, v němž se zaměřila na hlavní prvky definice obsažené v čl. 2 písm. a) směrnice 95/46/ES, včetně její části ohledně „identifikované nebo identifikovatelné“ osoby. Pracovní skupina v této souvislosti rovněž dospěla k závěru, že „anonymizované údaje jsou anonymní údaje, které se dříve týkaly identifikovatelné osoby, u nichž však již tato identifikace není možná“.

Pracovní skupina již tedy vysvětlila, že směrnice navrhuje používat jako kritérium pro posouzení, zda je proces anonymizace dostatečně spolehlivý, tj. zda je identifikace „přiměřeně“ nemožná, test „prostředků … které mohou být rozumně použity“. Na identifikovatelnost mají přímý vliv konkrétní souvislosti a okolnosti. V technické příloze k tomuto stanovisku je uvedena analýza dopadu nejvhodnější zvolené techniky.

Jak již bylo zdůrazněno, výzkum, nástroje a možnosti výpočetní techniky se vyvíjejí. Proto není možné ani užitečné uvádět vyčerpávající výčet okolností, za nichž již identifikace není možná. Některé důležité faktory si však zaslouží, aby byly zváženy a názorně doloženy.

Zaprvé lze tvrdit, že správci údajů by se měli soustředit na konkrétní prostředky, které by byly potřebné ke zrušení účinků technik anonymizace, zejména pokud jde o náklady a nezbytné know-how k zavedení těchto prostředků a o posouzení pravděpodobnosti jejich nasazení a jejich závažnosti. Mohli by například zvážit své úsilí a náklady v oblasti anonymizace (co se týče potřebného času i zdrojů) oproti stále větší dostupnosti nízkonákladových technických prostředků k identifikaci fyzických osob v souborech údajů, rostoucí dostupností dalších souborů údajů veřejnosti (například souborů zpřístupňovaných v rámci politik „přístupných údajů“) a mnoha příklady neúplné anonymizace, které mají negativní, a někdy i nenapravitelné, dopady na subjekty údajů[6]. Je třeba poznamenat, že nebezpečí identifikace se může časem zvyšovat a závisí i na vývoji informačních a komunikačních technologií. Případné právní předpisy tak musí být formulovány technologicky neutrálně a v ideálním případě zohledňovat změny v rozvíjejících se možnostech informačních technologií[7].

Zadruhé musí být „prostředky, které mohou být rozumně použity pro určení, zda je osoba identifikovatelná“, použity „správcem nebo jakoukoli jinou osobou“. Je tedy zásadní pochopit, že pokud správce údajů nevymaže původní (identifikovatelné) údaje na úrovni dané operace a část souboru údajů předá (například po odstranění nebo zamaskování identifikovatelných údajů), představuje výsledný soubor údajů i nadále osobní údaje. Výsledný soubor údajů lze kvalifikovat jako anonymní pouze v případě, že správce údajů sloučí údaje na úrovni, kde již nelze identifikovat jednotlivé operace. Například: jestliže organizace shromažďuje údaje o pohybech cestujících, budou jednotlivé vzorce cestování na úrovni operace u kteréhokoli účastníka stále představovat osobní údaje, neboť správce údajů (či jakákoli jiná osoba) má stále přístup k původním nezpracovaným údajům, i když byly ze souboru poskytovaného třetím stranám odstraněny přímé identifikátory. Pokud by však správce údajů vymazal nezpracované údaje a třetím stranám poskytoval pouze agregované statistiky na vyšší úrovni, například „v pondělí bývá na trase X o 160 % více cestujících než v úterý“, šlo by o anonymní údaje.

Účinný způsob anonymizace brání tomu, aby kdokoli mohl ze souboru údajů vyčlenit konkrétní osobu, propojit dva záznamy v souboru údajů (nebo ve dvou samostatných souborech údajů) a vyvozovat z takového souboru údajů jakékoli informace. Obecně vzato tedy odstranění přímo identifikujících prvků jako takových nestačí k tomu, aby již nebylo možno identifikovat subjekt údajů. Často bude nezbytné přijmout za účelem znemožnění identifikace další opatření, která budou opět záviset na souvislostech a zamýšlených cílech zpracovávání anonymizovaných údajů.

[6] Je zajímavé, že pozměňovací návrhy Evropského parlamentu k návrhu obecného nařízení o ochraně údajů, které byly předloženy v nedávné době (21. října 2013), uvádí konkrétně v 23. bodě odůvodnění, že: „Aby bylo možné zjistit, zda mohou být prostředky rozumně použity pro identifikaci fyzické osoby, je třeba přihlédnout ke všem objektivním faktorům, jako jsou náklady a doba potřebná pro identifikaci, přičemž je třeba vzít v úvahu jak technologii dostupnou v době zpracovávání údajů, tak technologický rozvoj“.

[7] Viz stanovisko pracovní skupiny zřízené podle článku 29 č. 4/2007, s. 15.

PŘÍKLAD:
Údaje o genetickém profilu jsou příkladem osobních údajů, které mohou být z důvodu jedinečné povahy určitých profilů ohroženy identifikací, pokud je jedinou použitou technikou odstranění totožnosti dárce. V literatuře již byly popsány situace[8], kdy kombinace veřejně přístupných genetických zdrojů (např. genealogické rejstříky, úmrtní oznámení, výsledky vyhledávání pomocí vyhledávačů) a metadata o dárcích DNA (termín darování, věk, bydliště) mohou odhalit totožnost některých osob, přestože DNA byla darována „anonymně“.

Obě skupiny technik anonymizace – randomizace a generalizace údajů[9] – mají své nedostatky. Každá z nich však může být za určitých okolností a souvislostí vhodná k dosažení kýženého cíle, aniž by bylo ohroženo soukromí subjektů údajů. Je třeba jasně říci, že „identifikací“ se rozumí nejen možnost odhalení jména a/nebo adresy fyzické osoby, ale součástí je také identifikovatelnost pomocí vyčlenění, možnost propojení a dedukce. Pro účely použití právního předpisu na ochranu údajů pak nezáleží na tom, jaké jsou úmysly správce údajů nebo jejich příjemců. Jakmile jsou údaje identifikovatelné, použijí se předpisy na ochranu údajů.

Pokud třetí strana zpracovává soubor údajů, u něhož byla použita technika anonymizace (byl anonymizován a zveřejněn původním správcem údajů), může tak legálně činit, aniž by brala v potaz požadavky na ochranu údajů, a to za předpokladu, že (přímo ani nepřímo) nemůže identifikovat subjekty údajů v původním souboru údajů. Třetí strany by však při rozhodování o tom, jak budou využívat, a především pak kombinovat takovéto anonymizované údaje pro své vlastní cíle, měly zohledňovat veškeré výše uvedené faktory související s okolnostmi a souvislostmi (včetně zvláštních rysů technik anonymizace tak, jak je použil původní správce údajů), neboť výsledné dopady pro ně mohou znamenat různé formy závazků. Pokud by z těchto faktorů a charakteristik mohlo vyplývat nepřijatelné riziko identifikace subjektů údajů, budou se na zpracování údajů opět vztahovat právní předpisy na ochranu údajů.

Výše uvedený seznam není v žádném případě vyčerpávající, poskytuje spíše obecný návod, jak přistupovat k posuzování identifikovatelnosti daného souboru údajů, na nějž byla uplatněna anonymizace v souladu s různými dostupnými technikami. Všechny výše uvedené aspekty lze považovat za mnohočetné rizikové faktory, které musí zvažovat jak správci údajů při anonymizaci souborů údajů, tak třetí strany při používání těchto „anonymizovaných“ souborů údajů pro vlastní účely.

2.2.3 Rizika používání anonymizovaných údajů

Při zvažování použití technik anonymizace musí správci údajů zohlednit následující rizika:

- Určitým nebezpečím je klást rovnítko mezi údaje anonymizované a pseudonymizované. V oddíle „Technická analýza“ bude vysvětleno, že pseudonymizované údaje nelze s anonymizovanými informacemi ztotožňovat, neboť pseudonymizované údaje i nadále umožňují vyčlenit konkrétní subjekty údajů a odhalit propojení mezi různými soubory údajů. Pseudonymita umožňuje identifikovatelnost, a proto spadá do oblasti působnosti právního režimu ochrany údajů. Zvláště důležité je to v souvislosti s vědeckým, statistickým nebo historickým výzkumem[10].

[8] Viz Bohannon, John, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, svazek 339, č. 6117, 18. ledna 2013, s. 262.
[9] Hlavní charakteristiky a rozdíly mezi těmito dvěma technikami anonymizace jsou popsány níže v oddílu 3 („Technická analýza“).
[10] Viz stanovisko pracovní skupiny zřízené podle článku 29 č. 4/2007, s. 18–20.

PŘÍKLAD:
Typickým příkladem mylných představ, které doprovázejí pseudonymizaci, je případ AOL (America On Line). V roce 2006 byla zveřejněna databáze obsahující dvacet milionů klíčových slov pro vyhledávání od 650 000 uživatelů za období tří měsíců, v níž byla jako jediné opatření pro zachování soukromí uživatelů provedena náhrada ID uživatelů AOL číselným znakem. Toto opatření vedlo k veřejné identifikaci a lokalizaci některých uživatelů. Jsou-li pseudonymizované řetězce dotazu ve vyhledávačích propojeny s jinými atributy, především například s IP adresami nebo s jinými konfiguračními parametry klienta, existuje velká pravděpodobnost identifikace.

- Další chybou je domnívat se, že náležitě anonymizované údaje (které splňují veškeré výše uvedené podmínky a kritéria a na něž se z jejich podstaty nevztahuje směrnice o ochraně údajů) zbavují fyzické osoby jakýchkoli záruk – především proto, že na používání těchto údajů se mohou vztahovat další právní předpisy. Například čl. 5 odst. 3 směrnice o ochraně soukromí v odvětví elektronických komunikací znemožňuje uchovávání „informací“ jakéhokoli druhu (včetně neosobních informací) v koncovém zařízení účastníka či uživatele nebo přístup k těmto informacím bez souhlasu účastníka či uživatele, neboť se jedná o součást obecnější zásady důvěrné povahy komunikací.

- Třetí chyba z nedbalosti plyne rovněž z toho, že není zvážen dopad, jaký mohou mít za určitých okolností náležitě anonymizované údaje na fyzické osoby, zejména v případě profilování. Oblast soukromého života osob je chráněna článkem 8 EÚLP a článkem 7 Listiny základních práv EU. Používání anonymizovaných souborů údajů zveřejněných pro účely použití třetími stranami může způsobit ztrátu soukromí, přestože na takový druh údajů již nelze použít právní předpisy na ochranu údajů. Zvláštní opatrnost při nakládání s anonymizovanými informacemi se vyžaduje především tehdy, používají-li se tyto informace (často v kombinaci s dalšími údaji) k přijímání rozhodnutí, která ovlivňují (byť nepřímo) fyzické osoby. Jak již bylo v tomto stanovisku zdůrazněno a jak objasnila pracovní skupina, zejména ve stanovisku k pojmu „účelového omezení“ (stanovisko č. 3/2013)[11], legitimní očekávání subjektů údajů týkající se dalšího zpracování jejich údajů by mělo být posuzováno z hlediska příslušných souvislostí, jako je povaha vztahu mezi subjekty údajů a správci údajů, platné právní závazky a transparentnost operací při zpracování.

3 Technická analýza, spolehlivost technologií a typické chyby

Existují různé postupy a techniky anonymizace, které jsou v různé míře spolehlivé. Tento oddíl bude věnován hlavním bodům, které musí správci údajů při používání těchto technik vzít v potaz, přičemž musí zohlednit zejména záruky, které dané techniky poskytují, s přihlédnutím k aktuálnímu stavu vývoje technologií a při zvážení tří rizikových oblastí, jež jsou nedílnou součástí anonymizace:

- *vyčlenění*, jež představuje možnost izolovat některé nebo všechny záznamy, které v daném souboru dat identifikují fyzickou osobu,
- *možnost propojení*, což je schopnost propojit nejméně dva záznamy týkající se téhož subjektu údajů nebo téže skupiny subjektů údajů (ve stejné databázi nebo ve dvou různých databázích). Může-li útočník zjistit (například pomocí srovnávací analýzy), že stejné skupině fyzických osob jsou přiděleny dva záznamy, ale nemůže z této skupiny vyčlenit jednotlivé osoby, je tato technika odolná proti „vyčlenění“, avšak nikoli proti možnosti propojení,

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

– *dedukce*, což je možnost vyvodit se značnou pravděpodobností hodnotu jednoho atributu z hodnot souboru jiných atributů.

Proti zpětné identifikaci by tak bylo spolehlivé řešení namířené proti těmto třem rizikům, jestliže by bylo prováděno nejpravděpodobnějšími a nejpřiměřenějšími prostředky, které může správce údajů a jakákoli třetí strana použít. Pracovní skupina v této souvislosti zdůrazňuje, že techniky znemožnění identifikace a anonymizace jsou předmětem probíhajícího výzkumu a že tento výzkum vytrvale ukazuje, že žádná technika sama o sobě není bez nedostatků. Obecně řečeno, existují dva odlišné přístupy k anonymizaci: první přístup je založen na ***randomizaci***, zatímco druhý na ***generalizaci***. Stanovisko se zabývá i pojmy, jako je *pseudonymizace, diferenciální utajení, l-diverzita a t-blízkost*.

Toto stanovisko používá v tomto oddílu následující terminologii: soubor údajů sestává z různých záznamů týkajících se fyzických osob (subjekty údajů). Každý záznam se týká jednoho subjektu údajů a skládá se ze souboru hodnot (nebo „vstupů“, např.: 2013) pro každý atribut (např. rok). Soubor údajů je soubor záznamů, který může mít buď podobu tabulky (či sady tabulek), nebo komentovaného/ohodnoceného grafu, což je dnes stále častější případ. Příklady uvedené ve stanovisku se budou týkat tabulek, ale lze je použít i na jiná grafická znázornění záznamů. Kombinaci atributů týkajících se subjektu údajů nebo skupiny subjektů údajů lze nazývat kvazi-identifikátory. V některých případech může být v souboru údajů více záznamů týkajících se téže fyzické osoby. „Útočník“ je třetí strana (tj. není ani správcem údajů, ani jejich zpracovatelem), která se náhodně nebo úmyslně dostane k původním záznamům.

3.1 Randomizace

Randomizace je skupina technik, kterými se mění věrohodnost údajů, čímž je odstraněna silná vazba mezi údaji a fyzickou osobou. Jsou-li údaje dostatečně nejisté, nemohou být nadále spojovány s konkrétní osobou. Randomizace samotná neubere žádnému záznamu na jedinečnosti, neboť každý záznam bude nadále odvozen od jediného subjektu údajů, ale může být ochranou před útokem nebo rizikem dedukce a v zájmu spolehlivějších záruk utajení ji lze kombinovat s technikami generalizace. Aby se zajistilo, že prostřednictvím záznamu nelze identifikovat konkrétní osobu, mohou být vyžadovány i další techniky.

3.1.1 Noise addition

Technika noise addition je užitečná především v případě, že atributy mohou mít významný negativní dopad na fyzické osoby, a spočívá v pozměnění atributů v souboru údajů tak, aby byly méně přesné, ale zachovaly si celkové rozdělení. Při zpracovávání souboru údajů bude mít pozorovatel dojem, že hodnoty jsou přesné, avšak to bude pravda jen do určité míry. Pokud byla například výška osoby původně změřena s přesností na nejbližší centimetr, anonymizovaný soubor údajů může obsahovat výšku s přesností jen +/–10 cm. Při účinném použití této techniky nebude třetí strana schopna ani identifikovat fyzickou osobu, ani opravit údaje nebo jinak zjistit, jak byly údaje pozměněny.

Noise addition je obvykle třeba kombinovat s dalšími technikami anonymizace, jako je odstranění zřejmých atributů a kvazi-identifikátorů. Míra nepřesnosti (noise) by měla záviset na požadované potřebě úrovně informací a na dopadu na soukromí fyzických osob v důsledku zveřejnění chráněných atributů.

3.1.1.1 Záruky

- *Vyčlenění:* Nadále je možné vyčlenit záznamy fyzické osoby (byť možná neidentifikovatelným způsobem), přestože záznamy jsou méně spolehlivé.
- *Možnost propojení:* Nadále je možné propojit záznamy jedné fyzické osoby, avšak záznamy jsou méně spolehlivé a skutečný záznam tak může být propojen s uměle přidaným záznamem (tj. se záznamem typu „noise“). V některých případech může chybné přiřazení vystavit subjekt údajů vysokému, ba dokonce vyššímu stupni rizika, než záznam správný.
- *Dedukce:* Útoky v podobě dedukce jsou možné, ale míra úspěšnosti bude nižší a některá falešná pozitiva (a falešná negativa) se mohou jevit věrohodná.

3.1.1.2 Obvyklé chyby

- Přidávání proměnlivých nepřesností (noise): Není-li nepřesnost sémanticky přijatelná (tj. pokud se vymyká měřítku a nerespektuje logiku atributů v souboru), může útočník, který má přístup k databázi, nepřesnost vyfiltrovat a v některých případech může obnovit chybějící vstupy. Jestliže je navíc soubor údajů příliš malý[12], může být i nadále možné propojit nepřesné údaje s externím zdrojem.
- Za jakého předpokladu je technika noise addition postačující: Noise addition je doplňkové opatření, které útočníkovi ztěžuje získání osobních údajů. Není-li míra nepřesnosti větší než informace obsažené v souboru údajů, nelze předpokládat, že by noise addition představovala samostatné a dostatečné řešení za účelem anonymizace.

3.1.1.3 Nedostatky metody noise addition

Velmi známým pokusem zahrnujícím zpětnou identifikaci byl pokus provedený s databází zákazníků poskytovatele videoobsahu Netflix. Výzkumní pracovníci analyzovali geometrické vlastnosti databáze sestávající z více než 100 milionů hodnocení více než 18 000 filmů vyjádřených téměř 500 000 uživateli na stupnici 1–5, kterou společnost zveřejnila poté, co byla „anonymizována“ v souladu s interní politikou ochrany soukromí a všechny informace, jež mohou identifikovat zákazníka, kromě hodnocení a dat, byly odstraněny. Byly přidány nepřesnosti a hodnocení byla mírně zvýšena nebo snížena.

Přesto bylo zjištěno, že 99 % záznamů uživatelů bylo možné v souboru údajů unikátně identifikovat za použití 8 hodnocení a dat se čtrnáctidenní chybou coby výběrových kritérií, přičemž omezení výběrových kritérií (2 hodnocení a třídenní chyba) stále umožňovalo identifikaci 68 % uživatelů[13].

3.1.2 Permutace

Tato metoda, která spočívá v přeházení hodnot atributů v tabulce tak, že aby některé z nich byly uměle propojeny s odlišnými subjekty údajů, je užitečná v případě, že je důležité zachovat rozdělení každého atributu v rámci souboru údajů.

Permutaci lze považovat za zvláštní formu noise addition. Při klasické technice noise addition jsou atributy pozměněny randomizovanými hodnotami. Vypracování soudržných nepřesností může být obtížné a pouze mírné pozměnění hodnot atributů nemusí přinést odpovídající

[12] Tato koncepce je dále rozpracována v příloze, s. 30.
[13] Narayanan, A. a Shmatikov, V., Robust de-anonymization of large sparse datasets, in: Security and Privacy, květen 2008, SP 2008, Sympozium IEEE o bezpečnosti a soukromí, s. 111–125.

utajení. Techniky permutace jsou alternativou, jíž se pozměňují hodnoty v rámci souboru údajů jen tak, že se u záznamů vymění. Toto přehození zajistí, že rozsah a rozmístění hodnot zůstanou stejné, ale změní se vzájemný vztah mezi hodnotami a jednotlivci. Existuje-li mezi dvěma atributy logický vztah nebo statistická korelace a jsou-li nezávisle vyměněny, tento vztah se zruší. Proto může být důležité přeházet soubor vzájemně souvisejících atributů tak, aby nebyl porušen logický vztah. Jinak by mohl útočník zjistit, že byly atributy vyměněny, a výměnu vrátit zpět.

Vezmeme-li například v úvahu dílčí soubor atributů v lékařském souboru údajů, jako jsou „důvody pro hospitalizaci / symptomy / příslušné oddělení“, bude hodnoty ve většině případů spojovat silný logický vztah a výměna pouze jedné hodnoty tak bude možné odhalit a dokonce vrátit zpět.

Podobně jako noise addition nemůže permutace sama o sobě zajistit anonymizaci a měla by být vždy kombinována s odstraněním zřejmých atributů/kvazi-identifikátorů.

3.1.2.1 Záruky

- *Vyčlenění:* Stejně jako u noise addition je nadále možné vyčlenit záznamy fyzické osoby, ale záznamy jsou méně spolehlivé.
- *Možnost propojení:* Pokud se permutace týká atributů a kvazi-identifikátorů, může zabránit „správnému“ internímu i externímu propojení atributů se souborem údajů, avšak stále umožňuje „nesprávné“ propojení, neboť skutečný záznam může být spojen s odlišným subjektem údajů.
- *Dedukce:* Ze souboru údajů lze stále provést dedukci, zejména pokud atributy vzájemně souvisej, nebo pokud mezi nimi existuje silný vzájemný logický vztah. Nicméně vzhledem k tomu, že útočník neví, které atributy byly vyměněny, musí zvážit možnost, že se jeho dedukce zakládá na chybné hypotéze, a proto je možné provádět pouze pravděpodobnostní dedukci.

3.1.2.2 Obvyklé chyby

- Výběr nesprávného atributu: výměna atributů, které nejsou citlivé nebo rizikové, by k ochraně osobních údajů nijak významně nepřispěla. Pokud by totiž citlivé/rizikové atributy zůstaly spojeny s původním atributem, byl by útočník i nadále schopen získat citlivé informace o fyzických osobách.
- Náhodná výměna atributů: Mají-li dva atributy silnou vazbu, náhodná výměna atributů neposkytne spolehlivé záruky. Tato obvyklá chyba je názorně ukázána v tabulce 1.
- Předpoklad, že permutace je postačující: Stejně jako noise addition nezajišťuje ani permutace anonymitu sama o sobě a měla by být kombinována s dalšími technikami, například s odstraněním zřejmých atributů.

3.1.2.3 Nedostatky metody permutace

Tento příklad ukazuje, jak náhodná výměna atributů znamená slabé záruky utajení, pokud mezi dvěma různými atributy existují logická spojení. Poté, co byl proveden pokus o anonymizaci, je stále velmi snadné odvodit příjem každé fyzické osoby podle zaměstnání (a roku narození). Na základě přímého prostudování údajů lze například říci, že generální ředitel v tabulce se pravděpodobně narodil v roce 1957 a má nejvyšší plat, zatímco nezaměstnaný se narodil v roce 1964 a jeho příjem je nejnižší.

Rok	Pohlaví	Zaměstnání	Příjem (provedena permutace)
1957	M	inženýr	70 tis.
1957	M	generální ředitel	5 tis.
1957	M	nezaměstnaný	43 tis.
1964	M	inženýr	100 tis.
1964	M	manažer	45 tis.

Tabulka 1: Příklad neúčinné anonymizace provedené pomocí výměny vzájemně propojených atributů.

3.1.3 Diferenciální utajení

Diferenciální utajení[14] patří do skupiny technik randomizace, ale přístupem se liší: zatímco přidávání nepřesností přichází na řadu ještě před předpokládaným zveřejněním souboru, diferenciální utajení lze použít tehdy, když správce údajů vytváří anonymizované podoby souboru údajů a zároveň si ponechává kopii původních údajů. Tyto anonymizované podoby jsou obvykle vytvářeny pro konkrétní třetí stranu pomocí dílčího souboru dotazů. Dílčí soubor obsahuje některé náhodné nepřesnosti, které jsou přidávány ex post. Diferenciální utajení správci údajů ukazuje, kolik nepřesností musí přidat a v jaké podobě, aby dosáhl nezbytných záruk utajení[15]. V této souvislosti bude zvláště důležité soustavně sledovat (přinejmenším u každého nového dotazu) všechny možnosti identifikace jednotlivých osob v souboru, jenž je výsledkem dotazů. Je však třeba objasnit, že techniky diferenciálního utajení nezmění původní údaje, takže dokud existuje originál, správce údajů může ve výsledcích dotazů diferenciálního utajení s přihlédnutím ke všem prostředkům, které mohou být rozumně použity, identifikovat jednotlivé osoby. I tyto výsledky je třeba považovat za osobní údaje.

Výhodou přístupu založeného na diferenciálním utajení je skutečnost, že se nezveřejňuje jednotlivý soubor údajů, nýbrž soubory údajů jsou předávány oprávněným třetím stranám v odpověď na konkrétní dotaz. Správce údajů si na pomoc při kontrole může ponechat seznam všech dotazů a požadavků a zaručit tak, že se třetí strany nedostanou k údajům, k nimž nemají oprávnění. Za účelem další ochrany soukromí mohou být na dotaz rovněž uplatněny techniky anonymizace, včetně přidávání nepřesností nebo substituce. Nalezení správného interaktivního mechanismu mezi dotazem a odpovědí, který dokáže dostatečně přesně (jen s menšími nepřesnostmi) zodpovědět jakýkoli dotaz a zároveň zachovat soukromí, zůstává otevřeným výzkumným problémem.

V zájmu omezení útoků na základě dedukce a možnosti propojení je nezbytné udržovat přehled o dotazech položených jedním subjektem a sledovat informace získané o subjektech údajů. Databáze na základě „diferenciálního utajení“ by tak neměly být uvolňovány pro veřejně přístupné vyhledávače, které nenabízejí vysledovatelnost dotazujících se subjektů.

3.1.3.1 Záruky

- *Vyčlenění:* Pokud jsou výstupem pouze statistiky a pravidla uplatněná na soubor jsou dobře zvolena, nemělo by být možné použít odpovědi k vyčlenění konkrétní osoby.

[14] Dwork, C., Differential privacy, in: Automata, languages and programming, Springer Berlin Heidelberg, 2006, s. 1–12).
[15] Srv. Felten, Ed, Protecting privacy by adding noise, 2012, dostupné na adrese: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/

- *Možnost propojení:* Při použití vícenásobných požadavků by mohlo být možné propojit ve dvou odpovědích údaje týkající se konkrétní osoby.
- *Dedukce:* Při použití vícenásobných požadavků je možné vyvodit informace o jednotlivých osobách nebo skupinách osob.

3.1.3.2 Obvyklé chyby

- Přidávání nedostatečného množství nepřesností: Aby se zabránilo propojování se základními poznatky, je důležité poskytnout minimální důkazy o tom, zda do souboru údajů přispěl konkrétní subjekt údajů nebo skupina subjektů údajů. Největším problémem z hlediska ochrany údajů je schopnost vytvořit dostatečné množství nepřesností, které se přidají k pravdivým odpovědím, aby tak bylo chráněno soukromí fyzických osob a přitom zůstala zachována užitečnost zveřejněných odpovědí.

3.1.3.3 Nedostatky metody diferenciálního utajení

Nakládání s každým dotazem nezávisle na ostatních: Kombinace výsledků dotazů může umožnit odhalení informací, které měly být tajné. Není-li uchovávána historie dotazů, může útočník do databáze „diferenciálního utajení“ naprojektovat řadu otázek, jimiž postupně zmenší rozptyl výsledného vzorku, v důsledku čehož se mohou s určitostí nebo s velmi vysokou pravděpodobností vyjevit konkrétní charakteristiky jednotlivého subjektu údajů nebo skupiny subjektů údajů. Dále je třeba upozornit, že je nutné vyhnout se mylnému přesvědčení, že údaje jsou pro třetí stranu anonymní, zatímco správce údajů je s přihlédnutím ke všem prostředkům, které mohou být rozumně použity, nadále schopen identifikovat subjekt údajů v původní databázi.

3.2 Generalizace

Generalizace je druhou skupinou technik anonymizace. Tento přístup spočívá v generalizování nebo naředění atributů subjektů údajů tak, že se pozmění příslušné měřítko nebo řád (tj. region místo města, měsíc místo týdne). Ačkoli generalizace může účinně zabránit vyčlenění, neumožňuje ve všech případech efektivní anonymizaci. Vyžaduje především specifický a propracovaný kvantitativní přístup, aby se vyloučila možnost propojení a dedukce.

3.2.1 Agregace a k-anonymita

Cílem technik agregace a k-anonymity je zabránit vyčlenění subjektu údajů tím, že je přiřazen ke skupině nejméně *k* dalších osob. K tomu je třeba generalizovat hodnoty atributů do takové míry, aby každá osoba sdílela stejné hodnoty. Sníží-li se například úroveň podrobnosti místa z města na zemi, je zahrnut větší počet subjektů údajů. Jednotlivá data narození mohou být generalizována na rozmezí dat nebo seskupena podle měsíce či roku. Další číselné atributy (např. platy, hmotnost, výška nebo dávka léčiva) lze generalizovat podle intervalů hodnot (např. plat 20 000 € – 30 000 €). Tyto metody lze použít, jestliže by vzájemný vztah přesných hodnot atributů mohl vytvořit kvazi-identifikátory.

3.2.1.1 Záruky

- *Vyčlenění:* Vzhledem k tomu, že *k* uživatelů nyní sdílí stejné atributy, by již nemělo být možné vyčlenit ze skupiny *k* uživatelů jednotlivé osoby.
- *Možnost propojení:* Ačkoli je možnost propojení omezená, je i nadále možné propojit záznamy podle skupin *k* uživatelů. V této skupině je pak pravděpodobnost, že dva

záznamy odpovídají stejným pseudoidentifikátorům, 1/k (a ta může být podstatně vyšší než pravděpodobnost, že takové záznamy nelze propojit).

- *Dedukce:* Hlavním nedostatkem modelu k-anonymity je skutečnost, že tato metoda nezabrání žádné formě útoku na základě dedukce. Nacházejí-li se totiž všechny *k* osoby v jedné skupině, je velmi snadné zjistit hodnotu této vlastnosti, je-li známo, do které skupiny osoba patří.

3.2.1.2 Obvyklé chyby

- Některé chybějící kvazi-identifikátory: Hlavním parametrem při posuzování k-anonymity je hranice *k*. Čím je hodnota *k* vyšší, tím silnější jsou záruky utajení. Obvyklou chybou je umělé navyšování hodnoty *k* zmenšováním zvažovaného souboru kvazi-identifikátorů. Omezování kvazi-identifikátorů usnadňuje vytvoření klastrů *k* uživatelů díky nedílné možnosti identifikace spojené s dalšími atributy (zejména jsou-li některé z nich citlivé nebo mají-li vysokou entropii, jako v případě velmi zřídkavých atributů). Hlavní chybou je, když se při výběru atributů ke generalizaci nepřihlíží ke všem kvazi-identifikátorům. Lze-li některé atributy použít k vyčlenění konkrétní osoby z klastru *k*, pak generalizace některé fyzické osoby nechrání (viz příklad v tabulce 2).
- Nízká hodnota *k*: Podobně problematická je snaha o nízkou hodnotu *k*. Je-li *k* příliš malé, je váha každé jednotlivé osoby v klastru příliš velká a útoky na základě dedukce mají vyšší míru úspěšnosti. Pokud například k = 2, je pravděpodobnost, že tyto dvě osoby mají stejné vlastnosti, vyšší než v případě, že k > 10.
- Vytváření skupin z osob, které nemají stejnou váhu: Problematické může být i vytváření skupin z osob s nestejným rozložením atributů. Dopad záznamu jednotlivé osoby na soubor údajů bude proměnlivý: některé osoby budou pro záznamy představovat významnou část, zatímco příspěvek jiných bude celkem bezvýznamný. Proto je důležité zajistit, aby *k* bylo dostatečně vysoké a žádní jedinci tak nepředstavovali příliš významnou část záznamů v klastru.

3.1.3.3 Nedostatky metody k-anonymity

Hlavním problémem k-anonymity je skutečnost, že nechrání před útoky na základě dedukce. Pokud v následujícím příkladu útočník ví, že se v souboru údajů nachází konkrétní osoba, která se narodila v roce 1964, ví rovněž, že tato osoba měla infarkt. Pokud dále víme, že soubor údajů byl získán od francouzské organizace, mají všechny osoby bydliště v Paříži, neboť pařížská poštovní směrovací čísla začínají číslicemi 750*.

Rok	Pohlaví	PSČ	Diagnóza
1957	M	750*	infarkt
1957	M	750*	cholesterol
1957	M	750*	cholesterol
1964	M	750*	infarkt
1964	M	750*	infarkt

Tabulka 2: Příklad špatně promyšlené k-anonymizace.

3.2.2 L-diverzita/t-blízkost

L-diverzita rozšiřuje k-anonymitu tak, aby se zajistilo, že útoky na základě deterministické dedukce již nebudou možné, neboť je zabezpečeno, aby ve všech ekvivalenčních třídách měl každý atribut nejméně *l* různých hodnot.

Základním cílem, jehož má být dosaženo, je omezit výskyt ekvivalenčních tříd s malou proměnlivostí atributů tak, aby útočníkovi, který má základní znalost konkrétního subjektu údajů, vždy zůstala značná nejistota.

L-diverzita je užitečná pro ochranu údajů před útoky na základě dedukce, kdy jsou hodnoty atributů dobře rozloženy. Je však třeba zdůraznit, že tato technika nemůže zabránit únikům informací, pokud jsou atributy uvnitř sekce rozloženy nerovnoměrně nebo spadají-li do malé škály hodnot nebo sémantických významů. L-diverzita v důsledku toho bývá předmětem útoků na základě pravděpodobnostní dedukce.

T-blízkost je zdokonalenou formou l-diverzity a jako taková má za cíl vytvořit ekvivalenční třídy, které napodobují původní rozložení atributů v tabulce. Tato technika je užitečná tam, kde je důležité, aby se údaje co nejvíce blížily původním údajům. Za tím účelem se na ekvivalenční třídu vztahuje další omezení, tedy nejen že by v každé ekvivalenční třídě mělo existovat *l* odlišných hodnot, ale každá hodnota by měla být zastoupena tolikrát, kolikrát je potřeba, aby to odráželo původní rozložení jednotlivých atributů.

3.2.2.1 Záruky

- *Vyčlenění:* Stejně jako k-anonymita mohou l-diverzita a t-blízkost zajistit, že v databázi nelze vyčlenit záznamy týkající se jedné osoby.
- *Možnost propojení:* Co se týče možnosti propojení, l-diverzita a t-blízkost nepředstavují lepší řešení než k-anonymita. Jedná se o tentýž problém jako s každým klastrem: pravděpodobnost, že stejné záznamy náleží stejnému subjektu údajů, je vyšší, než 1/N (kde N je počet subjektů údajů v databázi).
- *Dedukce:* L-diverzita a t-blízkost jsou lepším řešením než k-anonymita, neboť proti „l-diverzifikované“ a „t-blízké“ databázi již není možný útok na základě dedukce se 100% jistotou.

3.2.2.2 Obvyklé chyby

- Ochrana hodnot citlivých atributů jejich smíšením s jinými citlivými atributy: Pro poskytnutí záruk utajení nestačí mít dvě hodnoty atributu v klastru. Rozložení citlivých hodnot v každém klastru by se ve skutečnosti mělo podobat rozložení těchto hodnot v celém souboru, nebo by mělo být alespoň rovnoměrné v celém klastru.

3.2.2.3 Nedostatky metody l-diverzity

V níže uvedené tabulce je l-diverzita zaručena s ohledem na atribut „Diagnóza“. Víme-li však, že se v této tabulce vyskytuje osoba narozená v roce 1964, lze s vysokou pravděpodobností předpokládat, že tato osoba měla infarkt.

Rok	Pohlaví	PSČ	Diagnóza
1957	M	750*	**infarkt**
1957	M	750*	**cholesterol**
1957	M	750*	**cholesterol**
1957	M	750*	**cholesterol**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**cholesterol**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**
1964	M	750*	**infarkt**

Tabulka 3: Tabulka l-diverzity, kde hodnoty „Diagnóza“ nejsou rovnoměrně rozloženy.

Jméno	Datum narození	Pohlaví
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Tabulka 4: Ví-li útočník, že tyto osoby jsou uvedeny v tabulce 3, mohl by vyvodit, že měly infarkt.

4. Pseudonymizace

Pseudonymizace spočívá v nahrazení atributu v záznamu (obvykle jedinečného atributu) jiným atributem. Fyzickou osobu tak lze i nadále nepřímo identifikovat. Pokud se tedy použije pouze pseudonymizace, výsledkem nebude anonymní soubor údajů. Toto stanovisko se přesto věnuje i pseudonymizaci, neboť její používání doprovází řada chybných představ a omylů.

Pseudonymizace omezuje možnost propojení souboru údajů s původní totožností subjektu údajů. Jako taková je užitečným bezpečnostním opatřením, není však metodou anonymizace.

Výsledek pseudonymizace může být nezávislý na výchozí hodnotě (jako je tomu v případě náhodného počtu vygenerovaného správcem nebo příjmení zvoleného subjektem údajů), nebo může být odvozen od původních hodnot atributu či souboru atributů, např. pomocí hašování nebo systémem šifrování.

Nejpoužívanějšími technikami pseudonymizace jsou:

- Šifrování s tajným klíčem: V tomto případě může držitel klíče snadno zpětně identifikovat každý subjekt údajů, dešifruje-li soubor údajů, neboť osobní údaje zůstávají součástí soboru údajů, byť v zašifrované podobě. Předpokládáme-li, že byl použit nejmodernější systém šifrování, je dešifrování možné pouze se znalostí klíče.
- Hašování: Hašování je funkce, která vrací neměnnou velikost výstupu ze vstupu jakékoli velikosti (vstup může být jediným atributem nebo souborem atributů) a nelze ji vrátit zpět. To znamená, že již nepřetrvává nebezpečí vrácení, které existuje u šifrování. Pokud je však známa škála vstupních hodnot upravených hašováním, mohou být prostřednictvím hašování přehrány a u každého záznamu tak může být odvozena správná hodnota. Byl-li například soubor údajů pseudonymizován pomocí hašování národního identifikačního čísla, může být toto číslo jednoduše odvozeno hašováním všech možných vstupních hodnot a porovnáním výsledku s hodnotami v souboru údajů. Funkce hašování jsou obvykle koncipovány tak, aby mohly být rychle vypočítány, a bývají předmětem násilných útoků za použití hrubé síly[16]. Lze vytvořit i předem vypočítané tabulky, které umožňují hromadné vrácení velkého souboru hašovaných údajů.

  Použití funkce tzv. soleného hašování – salted hash (kde se k hašovanému atributu přidává náhodná hodnota nazývaná sůl) může snížit pravděpodobnost odvození vstupní hodnoty, avšak výpočet původní hodnoty atributu skryté za výsledkem hašování může být při použití rozumných prostředků přesto proveditelný[17].
- Zaklíčované hašování s uloženým klíčem: Jedná se o zvláštní hašovací funkci, která jako doplňkový vstup využívá tajný klíč (na rozdíl od soleného hašování, neboť „sůl“ obvykle není tajná). Správce údajů může přehrát funkci u atributu za použití tajného klíče, ale pro útočníka bez znalosti klíče je přehrát funkci mnohem obtížnější, neboť počet možností, které je třeba vyzkoušet, je dostatečně velký, aby tato technika byla nepraktická.

[16] Podstatou těchto útoků je vyzkoušení všech věrohodných vstupů, aby mohly být vytvořeny srovnávací tabulky.

[17] Zejména pokud je znám typ atributu (jméno, číslo sociálního pojištění, datum narození atd.). Pro přidání výpočetního požadavku se lze spolehnout na derivační hašování s klíčem, kdy je vypočtená hodnota několikrát hašována pomocí tzv. krátké soli (short salt).

- Deterministické šifrování nebo funkce zaklíčovaného hašování s vymazáním klíče: Tuto techniku lze srovnat s výběrem náhodného čísla coby pseudonymu pro každý atribut v databázi a následným vymazáním srovnávací tabulky. Toto řešení[18] umožňuje snížit riziko propojení mezi osobními údaji v souboru údajů a údaji týkajícími se téže osoby v jiném souboru údajů, kde je použit jiný pseudonym. Vezmeme-li v úvahu nejmodernější algoritmus, bude pro útočníka obtížné dešifrovat nebo přehrát funkci pomocí počítače, neboť by to za předpokladu, že klíč není k dispozici, znamenalo testovat všechny možné klíče.
- Tokenizace: Tato technika se obvykle používá (byť nikoli výlučně) ve finančnictví k nahrazení identifikačních čísel hodnotami, které pro útočníka snižují užitečnost. Je odvozena od předchozích funkcí a obvykle se zakládá na použití jednosměrných šifrovacích mechanismů nebo na přidělení sériového čísla či náhodně generovaného čísla, které není matematicky odvozeno od původního údaje, pomocí funkce indexace.

4.1 Záruky

- *Vyčlenění:* Nadále je možné vyčlenit záznamy jednotlivých osob, neboť osoby zůstávají identifikovány jedinečným atributem, který je výsledkem funkce pseudonymizace (= pseudonymizovaným atributem).
- *Možnost propojení:* Možnost propojení mezi záznamy používajícími stejný pseudonymizovaný atribut ve vztahu k téže osobě zůstává velmi snadná. I když jsou pro stejný subjekt údajů použity různé pseudonymizované atributy, propojení je nadále možné prostřednictvím jiných atributů. Pouze pokud nelze v souboru údajů použít k identifikaci subjektu údajů žádný jiný atribut a bylo-li odstraněno veškeré spojení mezi původním atributem a pseudonymizovaným atributem (včetně vymazání původních údajů), nebude mezi dvěma soubory údajů při použití rozdílných pseudonymizovaných atributů žádný zřejmý křížový odkaz.
- *Dedukce:* Útoky na skutečnou identitu subjektu údajů na základě dedukce jsou možné v rámci souboru údajů nebo napříč různými databázemi, které používají pro danou osobu stejný pseudonymizovaný atribut, nebo pokud jsou pseudonymy zcela zřejmé a náležitě nezakrývají identitu subjektu údajů.

4.2 Obvyklé chyby

- Přesvědčení, že pseudonymizovaný soubor údajů je anonymizovaný: Správci údajů se často domnívají, že odstranění nebo nahrazení jednoho nebo více atributů postačí k anonymizaci souboru údajů. Na mnoha příkladech bylo dokázáno, že tomu tak není. Pouhá změna identifikačního čísla nikomu nezabrání, aby identifikoval subjekt údajů, pokud v souboru údajů zůstanou kvazi-identifikátory, nebo jestliže mohou hodnoty ostatních identifikátorů stále odhalit totožnost fyzické osoby. V některých případech je identifikace osob v pseudonymizovaném souboru údajů stejně snadná jako s původními údaji. Aby byl soubor považován za anonymizovaný, je třeba provést další kroky, mezi něž patří odstranění a generalizace atributů či vymazání původních údajů, nebo alespoň vysoká míra agregace.
- Obvyklé chyby při použití pseudonymizace jako techniky k omezení možnosti propojení:

[18] V závislosti na dalších atributech v souboru údajů a na vymazání původních údajů.

- Použití stejného klíče v různých databázích: Odstranění možnosti propojení různých souborů údajů výrazně závisí na použití klíčovaného algoritmu a na tom, zda jednotlivá osoba bude v různých souvislostech odpovídat různým pseudonymizovaným atributům. Aby byla omezena možnost propojení, je důležité vyhnout se používání stejného klíče v různých databázích.
- Použití různých klíčů („rotujících klíčů“) pro různé uživatele: Může být lákavé používat různé klíče pro různé soubory uživatelů a klíče podle užití měnit (například použít stejný klíč k zápisu deseti vstupů týkajících se téhož uživatele). Pokud však tato operace není náležitě promyšlena, může způsobit výskyt vzorců, který částečně oslabuje zamýšlený užitek. Například střídání klíčů na základě zvláštních pravidel pro konkrétní osoby by usnadnilo možnost propojení vstupů odpovídajících daným fyzickým osobám. Stejně tak zmizení opakujících se pseudonymizovaných údajů z databáze v okamžiku, kdy se objeví nové údaje, může být znamením, že se oba záznamy týkají téže osoby.
- Zachování klíče: Je-li tajný klíč uchováván společně s pseudonymizovanými údaji a údaje jsou ohroženy, může útočník velmi snadno propojit pseudonymizované údaje s jejich původním atributem. Totéž platí, je-li klíč uchováván odděleně od údajů, ale není zabezpečen.

4.3 Nedostatky metody pseudonymizace

- *Zdravotnictví*

1. **Jméno, adresa, datum narození**	**2.** **Doba poskytování zvláštní pomoci**	**3.** **BMI**	**6.** **Referenční číslo výzkumné skupiny**
[redacted]	< 2 roky	15	QA5FRD4
	> 5 let	14	2B48HFG
	< 2 roky	16	RC3URPQ
	> 5 let	18	SD289K9
	< 2 roky	20	5E1FL7Q

Tabulka 5: Příklad anonymizace hašováním (jméno, adresa, datum narození), která může být snadno vrácena zpět.

Soubor údajů byl vytvořen za účelem prověření souvislosti mezi hmotností osoby a příjmem plateb zvláštní pomoci. Původní soubor údajů obsahoval jména, adresy a data narození subjektů údajů, ale tyto údaje byly vymazány. Referenční číslo výzkumné skupiny bylo generováno z vymazaných údajů za použití funkce hašování. Přestože byla jména, adresy a data narození z tabulky vymazána, jsou-li jméno, adresa a datum narození známy a je-li známa i použitá funkce hašování, je snadné vypočítat referenční čísla výzkumné skupiny.

- *Sociální sítě*

Bylo dokázáno[19], že z grafů týkajících se sociálních sítí lze navzdory použitým technikám anonymizace získat citlivé informace o konkrétních osobách. Provozovatel sociální sítě se mylně domníval, že pseudonymizace je natolik spolehlivá, aby neumožňovala identifikaci po prodeji údajů jiným společnostem pro účely marketingu a reklamy. Místo

[19] Narayanan, A. a Shmatikov, V., De-anonymizing social networks, 30. Sympozium IEEE o bezpečnosti a soukromí, 2009.

skutečných jmen použil provozovatel přezdívky, ale to zcela zjevně nestačilo k anonymizaci uživatelských profilů, neboť vztahy mezi různými osobami jsou jedinečné a mohou být použity jako identifikátor.

- *Umístění*

Výzkumní pracovníci z MIT[20] před nedávnem analyzovali pseudonymizovaný soubor údajů sestávající z informací o pohybech 1,5 milionu osob v prostoru a v čase na území v dosahu 100 km za dobu 15 měsíců. Prokázali, že 95 % osob lze vyčlenit na základě čtyř lokalizačních bodů a že pouhé dva body stačí na vyčlenění více než 50 % subjektů údajů (jeden z těchto bodů je znám a je jím s největší pravděpodobností „domov" nebo „kancelář"). Prostor pro ochranu soukromí je tudíž velmi omezený, ačkoli totožnost osob byla pseudonymizována tak, že jejich skutečné atributy [….] byly nahrazeny jinými štítky.

5. Závěry a doporučení

5.1 Závěry

Techniky znemožnění identifikace a anonymizace jsou předmětem intenzivního výzkumu a tento dokument v souladu s tím ukazuje, že každá technika má své výhody a nevýhody. V mnoha případech není možné dát minimální doporučení ohledně parametrů, které by měly být použity, neboť každý soubor údajů je třeba hodnotit samostatně.

V četných případech může anonymizovaný soubor údajů představovat pro subjekty údajů stále zbytkové riziko. Dokonce i v případě, kdy již nelze přesně dohledat záznam konkrétní osoby, může být stále možné shromáždit informace o této osobě pomocí jiných (veřejně či neveřejně) dostupných zdrojů informací. Je třeba zdůraznit, že kromě přímého dopadu důsledků špatně provedeného procesu anonymizace na subjekty údajů (zlost, ztráta času a pocit ztráty kontroly z toho, že osoba byla zahrnuta do klastru, aniž by o tom věděla nebo k tomu dala předchozí souhlas) se mohou objevit i další vedlejší důsledky špatné anonymizace, je-li subjekt údajů mylně zahrnut útočníkem do cílové skupiny jako výsledek zpracování anonymizovaných údajů, a to zejména v případě, že útočník má nedobré úmysly. Pracovní skupina tudíž zdůrazňuje, že techniky anonymizace mohou poskytnout záruky soukromí, ale pouze v případě, že jsou náležitě promyšleny, což znamená, že je nutné jasně stanovit nutné předpoklady (souvislosti) a cíl (cíle) procesu anonymizace, aby bylo dosaženo požadovaného stupně anonymizace.

5.2 Doporučení

- Některé techniky anonymizace vykazují nedílná omezení. Tato omezení musí být vážně vzata v úvahu dříve, než správci údajů danou techniku pro proces anonymizace použijí. Správci údajů musí zohlednit cíle, jichž má být anonymizací dosaženo, jako je například ochrana soukromí fyzických osob při zveřejnění souboru údajů nebo možnost získání informací ze souboru údajů.
- Žádná z technik popsaných v tomto dokumentu nesplňuje s jistotou kritéria účinné anonymizace (tj. nemožnost vyčlenění jednotlivé osoby, nemožnost propojení mezi záznamy týkajícími se jedné fyzické osoby a nemožnost dedukce týkající se dané osoby). Některým z těchto rizik však lze určitou technikou částečně nebo úplně zamezit, a proto je

[20] De Montjoye, Y.-A., Hidalgo, C., Verleysen, M. a Blondel, V., Unique in the Crowd: The privacy bounds of human mobility, Nature, č. 1376, 2013.

při navrhování použití určité techniky v konkrétní situaci a při uplatňování kombinace těchto technik coby způsobu posílení spolehlivosti výstupu důležité pečlivé plánování.

Níže uvedená tabulka poskytuje přehled silných a slabých stránek technik, jež jsou zvažovány ve vztahu ke všem třem základním požadavkům:

	Existuje stále riziko vyčlenění?	Existuje stále riziko možnosti propojení?	Existuje stále riziko dedukce?
Pseudonymizace	ano	ano	ano
Noise addition	ano	nemusí	nemusí
Náhrada	ano	ano	nemusí
Agregace nebo k-anonymita	ne	ano	ano
L-diverzita	ne	ano	nemusí
Diferenciální utajení	nemusí	nemusí	nemusí
Hašování/tokenizace	ano	ano	nemusí

Tabulka 6: Silné a slabé stránky zvažovaných technik.

- O optimálním řešení je třeba rozhodovat případ od případu. Řešení splňující všechna tři kritéria (tj. proces úplné anonymizace) by bylo spolehlivé proti identifikaci provedené prostředky, které může správce údajů nebo třetí osoba s největší pravděpodobností rozumně použít.
- Vždy, když návrh nesplňuje některé z kritérií, mělo by být provedeno pečlivé posouzení rizika identifikace. Pokud vnitrostátní právní předpisy vyžadují, aby příslušný orgán proces schválil nebo povolil anonymizaci, musí být posouzení tomuto orgánu předloženo.

Aby se snížilo riziko identifikace, mělo by být přihlédnuto k následujícím osvědčeným postupům:

Spolehlivé postupy anonymizace

Obecně:

- Nespoléhejte na postup „zveřejni a zapomeň“. Vzhledem ke zbytkovému riziku identifikace by správce údajů měl:
  - 1. pravidelně vyhodnocovat nová rizika a přehodnocovat zbytkové riziko (zbytková rizika);
  - 2. posoudit, zda jsou kontroly zaměřené na zjištění rizik dostatečné, a náležitým způsobem je upravit; a
  - 3. sledovat a kontrolovat rizika.
- Jakožto součást těchto zbytkových rizik berte v úvahu možnost identifikace neanonymizované části souboru údajů (pokud existuje), zejména v kombinaci s anonymizovanou částí, a popřípadě vzájemné vztahy mezi atributy (např. mezi zeměpisnými umístěními a údaji o majetku).

Kontextuální prvky:

- Měly by být jasně stanoveny cíle, jichž má být prostřednictvím anonymizace souboru údajů dosaženo, neboť hrají rozhodující úlohu při určování rizika identifikace.
- To úzce souvisí se zvážením všech příslušných kontextuálních prvků – např. povahy původních údajů, stávajících kontrolních mechanismů (včetně bezpečnostních opatření omezujících přístup k souborům údajů), velikosti vzorku (kvantitativní znaky), dostupnosti veřejných zdrojů informací (na něž se příjemci mohou spoléhat) a předpokládaného předání údajů třetím stranám (omezené, neomezené např. na internetu atd.).
- Měli by být vzati v potaz případní útočníci a mělo by se přihlížet k tomu, zda jsou údaje lákavé pro cílené útoky (v tomto ohledu budou rozhodujícími faktory opět citlivost informací a povaha údajů).

Technické prvky:

- Správci údajů by měli oznámit, jaké techniky anonymizace nebo jakou kombinaci technik anonymizace používají, zejména pokud chtějí zveřejnit anonymizovaný soubor údajů.
- Zjevné (např. zřídkavé) atributy / kvazi-identifikátory by měly být ze souboru údajů odstraněny.
- Pokud se používají techniky noise addition (při randomizaci), měla by být úroveň nepřesností přidávaných k záznamům stanovena jako funkce hodnoty atributu (tj. neměla by být přidávána žádná nepřesnost vymykající se měřítku), dopadu atributů, které mají být chráněny, na subjekty údajů a/nebo řídkosti souboru údajů.
- Pokud se (při randomizaci) spoléhá na diferenciální utajení, je třeba přihlédnout k potřebě sledovat dotazy a odhalit tak dotazy narušující soukromí, neboť rušivá povaha dotazů je kumulativní.
- Používají-li se techniky generalizace, je velmi důležité, aby se správci údajů neomezovali na jedno kritérium generalizace, byť i pro stejný atribut, tzn. je třeba vybrat různou úroveň podrobnosti míst nebo různé časové intervaly. Výběr použitého kritéria se musí řídit rozložením hodnot atributů v dané cílové skupině. Ne každé rozmístění je ke generalizaci vhodné – tzn. při generalizaci nelze uplatňovat univerzálnost. Je třeba zajistit variabilitu v rámci ekvivalenčních tříd. V závislosti na výše uvedených „kontextuálních prvcích“ (velikost vzorku atd.) by měla být zvolena konkrétní hranice, a pokud této hranice není dosaženo, měl by být konkrétní vzorek vyřazen (nebo by mělo být stanoveno odlišné kritérium generalizace).

PŘÍLOHA
Úvod do technik anonymizace

A.1. Úvod

Anonymita se v EU vykládá různě – v některých zemích odpovídá počítačové anonymitě (tj. i pro správce údajů ve spolupráci s jakoukoli jinou stranou by mělo být obtížné pomocí počítače přímo nebo nepřímo identifikovat jeden ze subjektů údajů) a v jiných zemích dokonalé anonymitě (tj. i pro správce údajů ve spolupráci s jakoukoli jinou stranou by mělo být nemožné pomocí počítače přímo nebo nepřímo identifikovat jeden ze subjektů údajů). V obou případech nicméně „anonymizace" znamená proces, při němž jsou údaje změněny na anonymní. Rozdíl spočívá v tom, co se považuje za přijatelnou míru rizika zpětné identifikace.

U anonymizovaných údajů lze předpokládat různorodé využití od sociálních průzkumů přes statistické analýzy po vývoj nových služeb či produktů. I takovéto činnosti s obecným zaměřením mohou mít někdy dopad na konkrétní subjekty údajů a porušit předpokládanou anonymní povahu zpracovaných údajů. Lze uvést řadu příkladů, od vedení cílených marketingových iniciativ až po uplatňování veřejných opatření založených na profilování uživatelů, na zvycích nebo na vzorcích mobility[21].

Kromě obecných prohlášení však bohužel neexistuje žádný rozvinutý systém měření, s jehož pomocí by bylo možné předem posoudit čas nebo úsilí potřebné ke zpětné identifikaci po zpracování nebo alternativně zvolit nejvhodnější postup, je-li třeba snížit pravděpodobnost, že se zveřejněná databáze bude týkat identifikovaného souboru subjektů údajů.

„Umění anonymizace", jak se tyto postupy někdy nazývají v odborné literatuře[22], je nové vědecké odvětví, které se dosud nachází v plenkách, a postupů, jak omezit možnosti identifikace souborů údajů, je velký počet. Je však třeba jasně uvést, že většina těchto postupů nebrání spojování zpracovaných údajů se subjekty údajů. Za určitých okolností byla anonymizace souborů údajů prokázána jako velmi úspěšná, v jiných situacích se projevil falešný optimismus.

Obecně vzato existují dva různé přístupy: jeden je založen na generalizaci atributů, druhý na randomizaci. Projdeme-li si podrobnosti a nuance těchto postupů, získáme nový náhled na možnosti identifikace údajů a v novém světle se nám vyjeví samotný koncept osobních údajů.

A.2. „Anonymizace" pomocí randomizace

Jednou z možností anonymizace je pozměnění stávajících hodnot tak, aby bylo znemožněno propojení mezi anonymizovanými údaji a původními hodnotami. Tohoto cíle lze dosáhnout pomocí celé řady metod, od přidávání nepřesností po záměnu údajů (permutaci). Je třeba zdůraznit, že odstranění atributu odpovídá extrémní podobě randomizace tohoto atributu (kdy je celý atribut překryt nepřesností).

Za určitých okolností není hlavním cílem celkového zpracování zveřejnění randomizovaného souboru údajů, ale umožnění přístupu k údajům pomocí dotazů. V tomto případě vyplývá riziko pro subjekt údajů z pravděpodobnosti, že útočník dokáže získat ze série různých dotazů

[21] Například případ TomTom v Nizozemsku (viz příklad objasněný v odstavci 2.2.3).
[22] Jun Gu, Yuexian Chen, Junning Fu, HuanchunPeng a Xiaojun Ye, Synthesizing: Art of Anonymization, Database and Expert Systems Applications, Lecture Notes in Computer Science, Springer, svazek 6261, 2010, s. 385–399.

informaci bez vědomí správce údajů. Aby byla osobám v souboru údajů zaručena anonymita, nemělo by být možné dojít k závěru, že subjekt údajů do souboru údajů přispěl, a měla by tak být přerušena vazba na jakékoli základní informace, které může útočník mít.

Riziko zpětné identifikace může dále snížit přidání nepřesnosti podle potřeby k odpovědi na dotaz. Tento přístup, v literatuře známý také jako diferenciální utajení[23], vychází z technik popsaných výše a na rozdíl od zpřístupnění veřejnosti umožňuje osobě, která údaje zveřejňuje, lépe kontrolovat přístup k údajům. Přidání nepřesnosti má dva základní cíle: jedním je ochrana soukromí subjektů údajů a druhým je zachování užitečnosti zveřejněných informací. Důležité je, že rozsah nepřesností musí být úměrný úrovni dotazů (příliš mnoho dotazů na jednotlivé osoby, které mají být zodpovězeny příliš přesně, zvyšuje pravděpodobnost identifikace). V současnosti je úspěšné použití randomizace třeba zvažovat případ od případu, přičemž žádná technika nenabízí jednoduchou metodu, neboť existují příklady úniků informací o atributech subjektu údajů (ať už obsažených v souboru údajů či nikoli) dokonce i v případech, kdy správce údajů považoval soubor údajů za randomizovaný.

Může být užitečné diskutovat o konkrétních příkladech a vyjasnit případné nedostatky randomizace coby prostředku k provedení anonymizace. V souvislosti s interaktivním přístupem mohou například dotazy považované za dotazy neohrožující soukromí znamenat pro subjekty údajů riziko. Pokud totiž útočník ví, že podskupina osob P je součástí souboru údajů, jenž obsahuje informace o výskytu atributu A v souboru S, může být pouhým položením dvou dotazů „Kolik osob v souboru S má atribut A?“ a „Kolik osob v souboru S, kromě těch, které patří do podskupiny P, má atribut A?“ možno (rozdílově) zjistit počet osob v podskupině P, které skutečně mají atribut A, a to jak deterministicky, tak na základě pravděpodobnostní dedukce. Soukromí osob v podskupině P může být v každém případě vážně ohroženo, zejména v závislosti na povaze atributu A.

V úvahu lze vzít rovněž skutečnost, že není-li subjekt údajů součástí souboru údajů, ale jeho vztah k údajům v souboru údajů je znám, zveřejnění souboru údajů může posléze ohrozit jeho soukromí. Je-li například známo, že „hodnota atributu A se u cíle liší kvantitou X od průměrné hodnoty v souboru“, může útočník pouze tím, že požádá správce databáze, aby provedl operaci nenarušující soukromí, a získá tak z databáze průměrnou hodnotu atributu A, přesně vyvodit osobní údaje týkající se konkrétního subjektu údajů.

Přidání určitých relativních nepřesností ke stávajícím hodnotám v databázi je operací, která musí být náležitě promyšlena. Pro účely ochrany soukromí je třeba přidat dostatečně velké nepřesnosti, které jsou však zároveň dostatečně malé, aby zůstala zachována užitečnost údajů. Pokud je například počet subjektů údajů se zvláštním atributem velmi malý nebo pokud je citlivost atributu vysoká, může být lepší uvést místo skutečného počtu rozmezí nebo obecnou větu, jako „malý počet případů, popřípadě dokonce nula“. Ačkoli je takto mechanismus zveřejnění s nepřesnostmi znám předem, soukromí subjektu údajů je zachováno, neboť nadále přetrvává určitý stupeň nejistoty. Pokud je nepřesnost náležitě promyšlena, jsou výsledky z hlediska užitečnosti stále využitelné pro statistické účely nebo pro účely rozhodování.

Randomizace databáze a přístup na základě diferenciálního utajení vyžadují další zamyšlení. Zaprvé, správná míra zkreslení se může výrazně lišit v závislosti na souvislostech (na typu dotazu, na počtu osob v databázi, na povaze atributu a jeho nedílné schopnosti identifikovat) a nelze počítat s řešením *ad omnia*. Souvislosti se navíc mohou měnit v čase a podle toho by se měl měnit také interaktivní mechanismus. Přesné vyměření nepřesností vyžaduje vysledování

[23] Dwork, Cynthia, Differential Privacy, Mezinárodní kolokvium o automatech, jazycích a programování (ICALP), 2006, s. 1–12.

kumulativního rizika pro soukromí, jemuž subjekty údajů vystavuje každý interaktivní mechanismus. Mechanismus přístupu k údajům pak musí být vybaven systémy varování pro situace, kdy je dosaženo „ceny za soukromí“ a kdy mohou být subjekty údajů vystaveny zvláštním rizikům, pokud je zodpovězen další dotaz. Cílem těchto systémů je pomoci správci údajů určit správnou míru zkreslení, kterou je pokaždé třeba uplatnit na konkrétní osobní údaje.

Na druhou stranu bychom měli uvažovat i o případech, kdy jsou hodnoty atributů vymazány (nebo pozměněny). Obvykle používaným řešením, jak se vypořádat s některými atypickými hodnotami pro atributy, je vymazání souboru údajů týkajících se netypických osob, nebo netypických hodnot. Ve druhém případě je pak důležité zajistit, aby se neexistence hodnot samotných nestala prvkem, podle něhož by bylo možno identifikovat subjekt údajů.

Nyní uvažujme o randomizaci na základě náhrady atributů. Velkým omylem týkajícím se anonymizace je její ztotožňování se šifrováním nebo kódováním. Východiskem tohoto omylu jsou dva předpoklady – především, že a) jakmile se na některé atributy záznamu v databázi (např. jméno, adresa, datum narození) použije šifrování, nebo pokud jsou tyto atributy nahrazeny zdánlivě randomizovanou řadou coby výsledkem klíčovaného kódování, například hašování s klíčem, je záznam „anonymizován“, a že b) anonymizace je účinnější, jestliže je délka klíče náležitá a šifrovací algoritmus je co nejmodernější. Tento omyl je mezi správci údajů velice rozšířený a vyžaduje objasnění, neboť se týká i pseudonymizace a jejích údajně nízkých rizik.

Zaprvé, cíle těchto technik jsou zásadně odlišné: cílem šifrování jakožto bezpečnostní techniky je důvěrná povaha komunikačního kanálu mezi identifikovanými stranami (lidmi, zařízeními nebo součástmi softwaru/hardwaru), aby se zabránilo odposlechu nebo k nechtěnému zveřejnění. Klíčované kódování odpovídá sémantické transformaci údajů v závislosti na tajném klíči. Cílem anonymizace je naopak znemožnit identifikaci osob tím, že se zabrání skrytému propojení atributů se subjektem údajů.

Ani šifrování, ani klíčované kódování se samo o sobě nehodí ke znemožnění identifikace subjektu údajů, neboť původní údaje jsou stále k dispozici v rukou správce údajů či je lze vydedukovat. Pouhé provedení sémantické transformace osobních údajů tak, jak je tomu u klíčovaného kódování, nevylučuje možnost obnovit původní strukturu údajů, a to buď pomocí opačného použití algoritmu, nebo útokem vedeným hrubou silou, v závislosti na povaze schémat nebo jako výsledek porušení údajů. Šifrování v souladu s nejnovějším vývojem může zajistit vyšší stupeň ochrany údajů, tj. pro subjekty, které neznají dešifrovací klíč, je nesrozumitelné, ale jeho výsledkem není nutně anonymizace. Dokud jsou klíč nebo původní údaje dostupné (a to i v případě důvěryhodné třetí strany, která je smluvně vázána, že poskytne službu úschovy pro klíč), není vyloučena možnost identifikace subjektu údajů.

Zaměřovat se pouze na spolehlivost mechanismu šifrování jakožto opatření zajišťujícího určitý stupeň „anonymizace“ souboru údajů je zavádějící, neboť celkovou bezpečnost mechanismu šifrování nebo hašování ovlivňuje mnoho dalších technických a organizačních faktorů. V literatuře je zaznamenáno mnoho úspěšných útoků, které algoritmus zcela obcházejí, a to tak, že využívají buď slabých stránek při opatrování klíče (např. existence méně bezpečného předem nastaveného módu), nebo jiných lidských faktorů (např. nedostatečně spolehlivých hesel při získávání klíče). Zvolené schéma šifrování s daným rozsahem klíče je určeno k zajištění důvěrné povahy po určenou dobu (většina běžných klíčů bude muset být okolo roku 2020 změněna), zatímco proces anonymizace by neměl být časově omezen.

Na tomto místě je vhodné podrobněji rozvést omezení randomizace atributů (nebo jejich nahrazování a odstraňování) s přihlédnutím k různým špatným příkladům anonymizace pomocí randomizace, k nimž došlo v posledních letech, a k důvodům těchto nedostatků.

Dobře známým příkladem špatně anonymizovaného souboru údajů je případ ceny Netflix[24]. Podíváme-li se na obecný záznam v databázi, v níž byla řada atributů týkajících se subjektu údajů randomizována, vidíme, že každý záznam lze následujícím způsobem přesto rozdělit na dva dílčí záznamy: {randomizované atributy, zřejmé atributy}, kde zřejmé atributy mohou být jakoukoli kombinací údajně neosobních údajů. Specifický postřeh, k němuž lze v souboru údajů ceny Netflix dospět, vychází z úvahy, že každý záznam může představovat bod ve vícerozměrném prostoru, kde každý zřejmý atribut tvoří souřadnici. Za použití této techniky lze na jakýkoli soubor údajů pohlížet jako na konstelaci bodů v tomto vícerozměrném prostoru, jež může vykazovat vysoký stupeň zřídkavosti, což znamená, že body od sebe mohou být vzdáleny. Ve skutečnosti mohou být vzdáleny natolik, že i po rozdělení prostoru na rozsáhlé oblasti bude každá oblast obsahovat pouze jeden záznam. Ani přidání nepřesností nepřiblíží záznamy dostatečně k sobě tak, aby se nacházely v téže vícerozměrné oblasti. Například v pokusu se službou Netflix byly záznamy dostatečně jedinečné, neboť obsahovaly pouze osm hodnocení filmů udělených za období 14 dnů. Po přidání nepřesností jak k hodnocení, tak k datům udělení nebylo možné pozorovat žádnou superpozici oblastí. Jinými slovy, stejný výběr pouhých osmi hodnocených filmů představoval otisk udělených hodnocení, která nebyla společná žádným dvěma subjektům údajů v databázi. Na základě tohoto geometrického pozorování spárovali výzkumní pracovníci údajně anonymní údaje služby Netflix s jinou veřejně přístupnou databází, v níž se hodnotí filmy (IMDB), a nalezli tak uživatele, kteří udělili hodnocení týmž filmům ve stejných časových rozmezích. Vzhledem k tomu, že většina uživatelů vykazovala jedinečnou shodu, mohly být pomocné informace nalezené v databázi IMDB importovány do zveřejněného souboru údajů služby Netflix, v němž tak byly údajně anonymizované záznamy obohaceny o totožnost uživatelů.

Důležité je zdůraznit, že se jedná o obecnou vlastnost: zbytková část každé „randomizované“ databáze vykazuje i nadále značné možnosti identifikace v závislosti na zřídkavosti kombinace zbytkových atributů. Jedná se o varovný aspekt, které by správci údajů měli mít při volbě randomizace coby způsobu cílené anonymizace neustále na mysli.

V mnoha pokusech tohoto druhu zabývajících se zpětnou identifikací byl použit obdobný přístup projekce dvou databází na stejnou dílčí oblast. Jde o velmi účinnou metodu zpětné identifikace, která se v posledních letech často používá v mnoha oblastech. Pokus s identifikací byl například proveden u sociální sítě[25], kde byl použit graf uživatelů pseudonymizovaný pomocí štítků. V tomto případě tvořil atributy použité k identifikaci seznam kontaktů jednotlivých uživatelů, neboť bylo prokázáno, že pravděpodobnost, že dvě osoby mají totožný seznam kontaktů, je velmi nízká. Na základě tohoto intuitivního předpokladu bylo zjištěno, že dílčí graf interních propojení mezi velmi omezeným počtem uzlů představuje topologický otisk, jejž je třeba získat a který je ukrytý v síti, a že jakmile se podaří tuto dílčí síť identifikovat, může být identifikována velká část celé sociální sítě. Pokud jde o některé údaje ohledně účinnosti podobného útoku, bylo prokázáno, že při použití méně než 10 uzlů (z nichž mohou vzniknout miliony různých dílčích síťových konfigurací, přičemž každá potenciálně představuje topologický otisk) může být sociální síť s více než 4 miliony

[24] Narayanan, Arvind, a Shmatikov, Vitaly, Robust De-anonymization of Large Sparse Datasets, Sympozium IEEE o bezpečnosti a soukromí, 2008, s. 111–125.
[25] Backstrom, L., Dwork, C., a Kleinberg, J. M., Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography, zápis z 16. Mezinárodní konference o celosvětové síti WWW'07, 2007, s. 181–190.

pseudonymizovaných uzlů a 70 miliony propojení náchylná k útokům, jejichž cílem je zpětná identifikace, a může být ohroženo utajení velkého množství spojení. Je třeba zdůraznit, že tento přístup ke zpětné identifikaci není zvlášť uzpůsoben konkrétním podmínkám sociálních sítí, je však dostatečně obecný, aby mohl být případně upraven pro jiné databáze, v nichž jsou zaznamenány vztahy mezi uživateli (např. telefonické kontakty, e-mailová korespondence, seznamky atd.).

Jiný způsob identifikace údajně anonymních záznamů se zakládá na analýze stylu psaní (stylometrie)[26]. K získání metrického systému z textu podrobeného rozboru byla již vyvinuta řada algoritmů, včetně četnosti používání určitých slov, výskytu zvláštních gramatických vzorců a druhu interpunkce. Všechny tyto charakteristické rysy lze použít k přiřazení údajně anonymního textu ke stylu psaní identifikovaného autora. Výzkumní pracovníci získali styl psaní z více než 100 000 blogů a v současnosti dokážou automaticky identifikovat autora příspěvku s přesností blížící se 80 %. Předpokládá se, že přesnost této techniky bude růst a využívat i další signály, například umístění nebo jiná metadata obsažená v textu.

Více pozornosti ze strany výzkumníků i celého oboru si zaslouží možnosti identifikace za použití sémantiky záznamu (tj. zbytkové nerandomizované části záznamu). Nedávné zpětné dohledání totožnosti dárců DNA (2013)[27] dokazuje, že od známého případu AOL (2006), kdy byla zveřejněna databáze obsahující 20 milionů klíčových slov od více než 650 000 uživatelů za období více než tří měsíců, došlo jen k velmi malému pokroku. Výsledkem byla identifikace a lokalizace řady uživatelů AOL.

Další skupinou údajů, která je zřídka anonymizována pouhým odstraněním totožnosti subjektů údajů nebo částečným zašifrováním některých atributů, jsou lokalizační údaje. Vzorce mobility osob mohou být dostatečně jedinečné, aby sémantická část lokalizačních údajů (místa, kde se daný subjekt v určitém okamžiku nacházel) mohla i bez dalších atributů odhalit mnohé rysy subjektu údajů[28]. To bylo mnohokrát prokázáno v reprezentativních akademických studiích[29].

V této souvislosti je třeba varovat před používáním pseudonymů coby způsobu odpovídající ochrany subjektů údajů před únikem totožnosti nebo atributů. Je-li pseudonymizace založena na náhradě totožnosti jiným jedinečným kódem, je předpoklad, že představuje spolehlivou ochranu před zpětnou identifikací, naivní a nebere v úvahu složitost metod identifikace a rozmanitost souvislostí, v nichž mohou být použity.

A.3. „Anonymizace“ pomocí generalizace

Jednoduchý příklad může pomoci vysvětlit přístup založený na generalizaci atributů.

---

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

[27] Genetické údaje jsou mimořádně významným příkladem citlivých údajů, které mohou být ohroženy zpětnou identifikací, pokud je jediným mechanismem jejich „anonymizace“ odstranění totožnosti dárců. Viz příklad uvedený výše v odstavci 2.2.2. Viz rovněž Bohannon, John, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, svazek 339, č. 6117, 18. ledna 2013, s. 262.

[28] Tomuto problému se věnují některé vnitrostátní právní předpisy. Například ve Francii jsou zveřejněné lokalizační statistiky anonymizovány pomocí generalizace a permutace. Institut INSEE tak zveřejňuje statistiky, které jsou generalizovány na základě agregace údajů na oblast 40 000 metrů čtverečních. Úroveň podrobnosti souboru údajů je dostatečná, aby se zachovala užitečnost údajů, a permutace chrání před útoky namířenými proti anonymizaci ve zřídkavých oblastech. Obecně lze říci, že agregace této skupiny údajů a jejich permutace poskytují silné záruky před útoky na základě dedukce a rušení anonymizace: http://www.insee.fr/en/

[29] De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. a Blondel, V. D., Unique in the Crowd: The privacy bounds of human mobility, Nature, č. 1376, 2013.

Uvažujme o případu, kdy se správce údajů rozhodne zveřejnit jednoduchou tabulku obsahující tři informace (nebo atributy): identifikační číslo, které je pro každý záznam jedinečné, identifikaci místa, která spojuje subjekt údajů s místem, kde žije, a informaci o majetku, jež ukazuje, jaký majetek subjekt údajů má. Dále předpokládejme, že tento majetek je jednou ze dvou odlišných hodnot, které jsou obecně označeny jako {P1, P2}:

Sériové identifikační číslo	**Identifikace místa**	**Majetek**
#1	Řím	P1
#2	Madrid	P1
#3	Londýn	P2
#4	Paříž	P1
#5	Barcelona	P1
#6	Milán	P2
#7	New York	P2
#8	Berlín	P1

Tabulka A1: Vzorek subjektů údajů seskupených podle místa a majetku P1 a P2.

Jestliže někdo, koho nazveme útočník, předem ví, že v tabulce se vyskytuje konkrétní subjekt údajů (cíl), který žije v Miláně, může se po prostudování tabulky dozvědět, že jelikož má tuto identifikaci místa pouze subjekt údajů č. 6, je rovněž vlastníkem majetku P2.

Tento naprosto základní příklad ilustruje hlavní prvky jakéhokoli postupu identifikace, které se použijí na soubor údajů, jenž byl podroben údajnému procesu anonymizace. Konkrétně se zde vyskytuje útočník, který (náhodou či záměrně) zná základní informace o některých nebo všech subjektech údajů v souboru údajů. Cílem útočníka je spojit tuto znalost souvislostí s údaji ve zveřejněném souboru údajů a získat tak jasnější obraz o charakteristických rysech těchto subjektů údajů.

Aby bylo spojení údajů s jakoukoli znalostí souvislostí méně efektivní nebo méně bezprostřední, měl by se správce údajů zaměřit na identifikaci místa a nahradit město, v němž subjekty údajů žijí, širší oblastí, například zemí. Tabulka by pak vypadala takto:

Sériové identifikační číslo	**Identifikace místa**	**Majetek**
#1	Itálie	P1
#2	Španělsko	P1
#3	Spojené království	P2
#4	Francie	P1
#5	Španělsko	P1
#6	Itálie	P2
#7	USA	P2
#8	Německo	P1

Tabulka A2: Generalizace tabulky A1 podle státní příslušnosti.

Při této nové agregaci údajů neumožňuje útočníkova znalost základních informací o identifikovaném subjektu údajů (tedy „cíl žije v Římě a je uveden v tabulce“) vytvoření žádného jasného závěru, pokud jde o jeho majetek. Je tomu tak proto, že dva Italové v tabulce mají rozdílný majetek P1 a P2. Útočník má padesátiprocentní nejistotu, pokud jde o majetek cílového subjektu. Tento jednoduchý příklad ukazuje účinek generalizace na systém anonymizace. Přestože by však tento trik s generalizací mohl být účinný, neboť snižuje pravděpodobnost identifikace italského cíle na polovinu, není účinný ve vztahu k cílům z jiných míst (např. USA).

Kromě toho může útočník i nadále zjistit informace o španělských cílech. Pokud má útočník základní znalosti typu „cíl žije v Madridu a je uveden v tabulce“ nebo „cíl žije v Barceloně a je uveden v tabulce“, může se stoprocentní jistotou vyvodit, že cíl vlastní majetek P1. Generalizace tudíž neposkytuje všem osobám v souboru údajů stejnou úroveň soukromí nebo odolnosti před útoky na základě dedukce.

Na základě této úvahy může být člověk v pokušení dospět k závěru, že by mohla být užitečná ještě větší generalizace, jež by mohla zabránit jakémukoli propojení – například generalizace podle kontinentu. Tabulka by pak vypadala takto:

Sériové identifikační číslo	**Identifikace místa**	**Majetek**
#1	Evropa	P1
#2	Evropa	P1
#3	Evropa	P2
#4	Evropa	P1
#5	Evropa	P1
#6	Evropa	P2
#7	Severní Amerika	P2
#8	Evropa	P1

Tabulka A3: Generalizace tabulky A1 podle kontinentu.

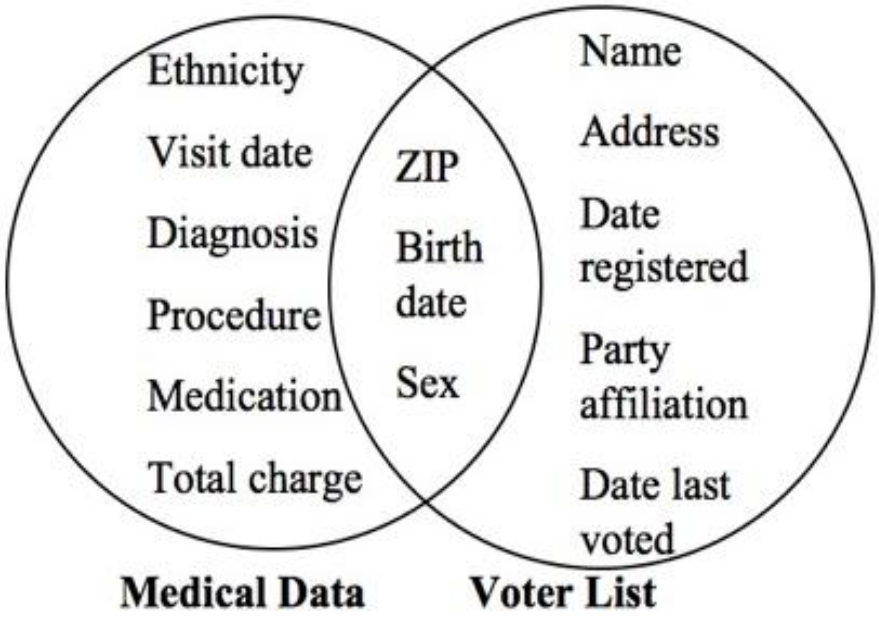
Při tomto druhu agregace by všechny subjekty údajů v tabulce, kromě subjektu žijícího v USA, byly chráněny před útoky na základě propojení a identifikace a veškeré informace o souvislostech, jako „cíl žije v Madridu a je uveden v tabulce“ nebo „cíl žije v Milánu a je uveden v tabulce“, by vedly pouze k určité míře pravděpodobnosti, pokud jde o majetek, který se vztahuje k danému subjektu údajů (P1 s pravděpodobností 71,4 % a P2 s pravděpodobností 28,6 %), nikoli k přímému propojení. Tato další generalizace však vede rovněž ke zřejmé a radikální ztrátě informací: tabulka neumožňuje odhalit případné vzájemné vztahy mezi majetkem a místem, konkrétně skutečnost, zda by konkrétní místo mohlo s větší pravděpodobností souviset s jedním ze dvou typů majetku, neboť uvádí pouze tzv. „marginální“ rozložení, tedy absolutní pravděpodobnost výskytu majetku P1 a P2 v celém souboru (v našem příkladu tedy 62,5 % a 37,5 %) a na obou kontinentech (jak bylo řečeno, 71,4 % a 28,6 % v Evropě a 100 % a 0 % v Severní Americe).

Z příkladu rovněž vyplývá, že technika generalizace má vliv na praktickou užitečnost údajů. V současnosti jsou k dispozici některé technické nástroje, s jejichž pomocí lze předem (tj. ještě před zveřejněním souboru údajů) určit, jaký je nejvhodnější stupeň generalizace atributů

tak, aby se omezila rizika identifikace subjektů údajů v tabulce, aniž by byla nepřiměřeně ovlivněna užitečnost zveřejněných údajů.

K-anonymita

Snaha zabránit útokům na základě propojení, jež vychází z generalizace atributů, je známa pod pojmem k-anonymita. Tato technika má původ v pokusu se zpětnou identifikací, který byl proveden koncem 90. let a v němž společnost z USA působící ve zdravotnictví zveřejnila údajně anonymizovaný soubor údajů. Tato anonymizace spočívala v odstranění jmen subjektů údajů, avšak soubor údajů nadále obsahoval údaje o zdravotním stavu a další atributy, jako je PSČ (identifikace místa, kde subjekty žijí), pohlaví a úplné datum narození. Stejná trojice údajů {PSČ, pohlaví, úplné datum narození} byla uvedena i v jiných veřejně přístupných rejstřících (např. v seznamu voličů) a akademický výzkumný pracovník ji tak mohl použít k propojení totožnosti určitých subjektů údajů s atributy zveřejněného souboru údajů. Útočník (výzkumný pracovník) mohl znát následující základní informace: „Vím, že subjekt údajů v seznamu voličů s konkrétní trojicí údajů {PSČ, pohlaví, úplné datum narození} je jedinečný. Záznam s touto trojicí údajů existuje ve zveřejněném souboru údajů“. Empiricky bylo zjištěno[30], že velkou většinu (více než 80 %) subjektů údajů ve veřejném rejstříku, který byl v tomto výzkumném pokusu použit, lze jednoznačně spojit s konkrétní trojicí údajů, která umožnila identifikaci. V tomto případě tedy údaje nebyly náležitě anonymizovány.



Obrázek A1. Zpětná identifikace na základě propojení údajů.

Aby se snížila účinnost podobných útoků založených na propojení, uvádí se, že správci údajů by měli nejdříve prozkoumat soubor údajů a seskupit ty atributy, které by útočník mohl přiměřeně použít za účelem propojení zveřejněné tabulky s jiným pomocným zdrojem. Každá skupina by měla obsahovat nejméně *k* identických kombinací generalizovaných atributů (tj. měla by představovat ekvivalenční třídu atributů). Soubory údajů by pak měly být zveřejňovány až poté, co budou rozděleny na takovéto homogenní skupiny. Atributy vybírané ke generalizaci jsou v literatuře známy jako kvazi-identifikátory, neboť jejich přímá znalost by vedla k okamžité identifikaci subjektů údajů.

[30] Sweeney, L., Weaving Technology and Policy Together to Maintain Confidentiality, Journal of Law, Medicine & Ethics, 25, č. 2&3, 1997, s. 98–110.

Řada pokusů s identifikací odhalila slabé stránky špatně sestavených k-anonymizovaných tabulek. Může se tak stát například tehdy, jestliže další atributy v ekvivalenční třídě jsou totožné (jako je tomu u ekvivalenční třídy španělských subjektů údajů v příkladu v tabulce 2), nebo je-li jejich rozložení velmi nerovnoměrné a převládají specifické atributy, anebo proto, že počet záznamů v ekvivalenční třídě je velmi malý, což v obou případech zvyšuje pravděpodobnost možné dedukce, či proto, že mezi zřejmými atributy ekvivalenčních tříd neexistuje žádný výrazný „sémantický“ rozdíl (např. kvantitativní míra těchto atributů může být skutečně rozdílná, ale číselně se značně přibližuje, nebo mohou náležet k řadě sémanticky podobných atributů, např. stejné rozmezí úvěrového rizika nebo stejná skupina chorob), takže ze souboru údajů může i nadále unikat velké množství informací o subjektech údajů, které lze použít při útocích na základě propojení[31]. Na tomto místě je důležité podotknout, že jakmile jsou údaje zřídkavé (například když se v dané zeměpisné oblasti jen málokdy vyskytuje určitý majetek) a při první agregaci není možné seskupit údaje s dostatečným množstvím výskytů různého typu majetku (v zeměpisné oblasti lze nalézt jen nízký výskyt několika málo typů majetku), je nezbytné provést další agregaci atributů, aby bylo dosaženo zamýšlené anonymizace.

L-diverzita

Na základě těchto postřehů byly v průběhu let navrženy různé varianty k-anonymity a byla vypracována některá technická kritéria pro posílení techniky anonymizace pomocí generalizace, jejichž cílem je snížení rizika útoků na základě propojení. Vycházejí z pravděpodobnostních vlastností souborů údajů. Konkrétně se přidává další omezení, jmenovitě to, že každý atribut v ekvivalenční třídě se vyskytuje nejméně *l*-krát, takže útočník má vždy ohledně atributů značnou nejistotu, a to i v případě, že zná základní informace o konkrétních subjektech údajů. Je to totéž, jako kdybychom řekli, že v souboru údajů (nebo jeho části) by měl být obsažen minimální počet výskytů zvolené vlastnosti: tento trik může snížit riziko zpětné identifikace. To je cílem techniky anonymizace pomocí l-diverzity. Příklad této techniky je uveden v tabulce A4 (původní údaje) a A5 (výsledek zpracování). Jak je zřejmé, díky řádnému zpracování identifikace místa a věku jednotlivých osob v tabulce A4 vede generalizace atributů k podstatnému zvýšení nejistoty, pokud jde o skutečné atributy každého subjektu údajů v průzkumu. I když útočník například ví, že subjekt údajů patří do první ekvivalenční třídy, nemůže dále zjistit, zda osoba vlastní majetek X, Y nebo Z, neboť v této třídě (a ve všech ostatních ekvivalenčních třídách) existuje přinejmenším jeden záznam vykazující tyto vlastnosti.

[31] Je třeba zdůraznit, že ačkoli byly záznamy s údaji seskupeny podle atributů, je možné zjistit vzájemné vztahy. Pokud správce údajů zná druhy vzájemných vztahů, které chce ověřit, může vybrat nejvhodnější atributy. Například výsledky průzkumů PEW nejsou předmětem podrobně propracovaných útoků na základě dedukce a jsou stále velmi užitečné při zjišťování vzájemných vztahů mezi demografickými údaji a zájmy: http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx

Sériové číslo	Identifikace místa	Věk	Majetek
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tabulka A4: Tabulka s jednotlivci seskupenými podle místa, věku a tří typů majetku X, Y a Z.

Sériové číslo	Identifikace místa	Věk	Majetek
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Tabulka A5: Příklad verze tabulky A4 zpracované pomocí l-diverzity.

T-blízkost

Zvláštní případ atributů v části souboru, které jsou rozloženy nerovnoměrně nebo náleží k malé škále hodnot či sémantických významů, se řeší přístupem známým pod názvem t-blízkost. Jedná se o další zdokonalení anonymizace pomocí generalizace a spočívá v uspořádání údajů tak, aby vznikly ekvivalenční třídy, které co nejvíce odráží počáteční rozložení atributů v původním souboru údajů. Pro tento účel se v zásadě používá postup o dvou krocích. Tabulka A6 je původní databáze, včetně zřejmých záznamů o subjektech údajů, které jsou seskupeny podle místa, věku, platu a dvou skupin sémanticky podobných vlastností, tedy (X1, X2, X3) a (Y1, Y2, Y3) (např. podobného úvěrového rizika či podobných nemocí). V tabulce byla v první fázi provedena l-diverzifikace, kde $l = 1$ (tabulka A7), a to tak, že záznamy byly seskupeny do sémanticky podobných ekvivalenčních tříd a byla provedena nedostatečná cílená anonymizace. Poté byla tabulka zpracována tak, aby bylo v každé části souboru dosaženo t-blízkosti (tabulka A8) a vyšší variability. Ve druhém kroku tak každá ekvivalenční třída obsahuje záznamy z obou majetkových skupin. Za povšimnutí stojí, že identifikace místa a věk mají v různých krocích postupu různou úroveň podrobnosti: to znamená, že každý atribut by mohl vyžadovat různá kritéria generalizace, aby bylo dosaženo cílené anonymizace, a od správců údajů to tím pádem vyžaduje zvláštní plánování a příslušnou výpočetní zátěž.

Sériové číslo	Identifikace místa	Věk	Plat	Majetek
1	1127	29	30 tis.	X1
2	1112	22	32 tis.	X2
3	1128	27	35 tis.	X3
4	1215	43	50 tis.	X2
5	1219	52	120 tis.	Y1
6	1216	47	60 tis.	Y2
7	1115	30	55 tis.	Y2
8	1123	36	100 tis.	Y3
9	1117	32	110 tis.	X3

Tabulka A6: Tabulka s osobami seskupenými podle místa, věku, platů a dvou skupin majetku.

Sériové číslo	Identifikace místa	Věk	Plat	Majetek
1	11**	2*	30 tis.	X1
2	11**	2*	32 tis.	X2
3	11**	2*	35 tis.	X3
4	121*	>40	50 tis.	X2
5	121*	>40	120 tis.	Y1
6	121*	>40	60 tis.	Y2
7	11**	3*	55 tis.	Y2
8	11**	3*	100 tis.	Y3
9	11**	3*	110 tis.	X3

Tabulka A7: Verze tabulky A6 zpracovaná pomocí l-diverzifikace.

Sériové číslo	Identifikace místa	Věk	Plat	Majetek
1	112*	<40	30 tis.	X1
3	112*	<40	35 tis.	X3
8	112*	<40	100 tis.	Y3
4	121*	>40	50 tis.	X2
5	121*	>40	120 tis.	Y1
6	121*	>40	60 tis.	Y2
2	111*	<40	32 tis.	X2
7	111*	<40	55 tis.	Y2
9	111*	<40	110 tis.	X3

Tabulka A8: Verze tabulky A6 zpracovaná pomocí t-blízkosti.

Je třeba jasně uvést, že cíle generalizace atributů subjektů údajů takovýmito erudovanými způsoby lze někdy dosáhnout pouze u malého počtu záznamů, nikoli u všech. Osvědčené postupy mohou zajistit, aby každá ekvivalenční třída obsahovala více jednotlivců a aby nebyl možný žádný útok na základě dedukce. Tento přístup každopádně vyžaduje, aby správci údajů provedli hloubkové posouzení dostupných údajů a kombinatorické hodnocení různých alternativ (například různé amplitudy rozmezí, různá místa nebo úroveň podrobnosti údajů o věku atd.). Jinými slovy, anonymizace pomocí generalizace nemůže být výsledkem prvního přibližného pokusu správců údajů nahradit analytické hodnoty atributů v záznamu rozmezím, neboť jsou zapotřebí specifičtější kvantitativní přístupy, jako je vyhodnocení entropie atributů v každé části souboru nebo změření rozdílu mezi původním rozložením údajů a rozložením v jednotlivých ekvivalenčních třídách.