Anonim módon adott meg adatokat? Akkor is azonosítható!

Azt hitte, hogy ha anonimizálják az adatait, máris biztonságban van? Egy kutatócsapat bebizonyította: ennél nagyobbat nem is tévedhetett volna.

Hiába anonimizálják az orvosi vagy népszámlálási adatokat, az egyéb jellemzők alapján közel százszázalékos pontossággal visszaállítható, hogy melyik adat név szerint kihez tartozik. Ez igencsak feladja a leckét a GDPR-felelősöknek.

Az európai általános adatvédelmi rendelet előírja, hogy az anonimizálást úgy kell elvégezni, hogy a tárolt adatok alapján a kapcsolat ne legyen többé helyreállítható egy természetes személy és a rá vonatkozó adat között. Már a rendelet hatályba lépésekor is sokan felhívták arra a figyelmet, hogy ezt nem is olyan egyszerű biztosítani. És mint kiderült, bizonyos esetekben ez szinte lehetetlen is az adatkészlet használhatatlanná tétele nélkül.

Az anonimizált adatokra a GDPR sem vonatkozik

A The New York Times számol be egy kutatásról, amely bizonyította: az Amerikai Népszámlálási Hivatal (U.S. Census Bureau) egyébként anonimizált adataiból simán vissza lehet állítani, hogy melyik amerikai polgár mit nyilatkozott a népszámlálást végző kérdezőbiztosoknál. A Nature Communicationsben publikált módszer, amit a londoni Imperial College és a Leuveni Katolikus Egyetem fiatal kutatói dolgoztak ki, szinte minden hasonló anonimizált adathalmazra (egészségügyi adatok, közvélemény-kutatási felmérések stb.) alkalmazható. Ez komoly visszaélésekre ad lehetőséget, hiszen például egészségügyi biztosítók, cégek, pártok vagy akár állami szervek juthatnak olyan információkhoz a polgárokról, amiket egyébként nem lenne joguk elkérni és tárolni.

A legtöbb országban az anonimizált adatokra nem vonatkoznak az adatvédelmi előírások. A GDPR is így rendelkezik, ezért például nem kell alkalmazni statisztikai vagy kutatási célú adatkezelésnél. Az ilyen adatokra nagy a kereslet, használják közvéleménykutatók, politikusok, vállalatok egyaránt, hogy megismerjék politikai, vallási, szexuális, vásárlási stb. preferenciáinkat – természetesen szigorúan statisztikai alapon.

Néhány jellemzőből megmondom, ki vagy

Csakhogy ezek a hatalmas adatkészletek jellemzően tartalmaznak minden benne szereplő, egyébként anonimizált személyről különböző egyedi jellemzőket, attribútumokat. A kutatók felhozzák példának az egyik amerikai adatbróker céget, amely olyan, egyébként anonimizált adatkészletet árult ügyfeleinek, amely 120 millió amerikairól tartalmazott háztartásonként 248 jellemzőt. Vagy a Cambrige-i Egyetem kutatói egy olyan – szintén anonimizált – adatkészletet osztottak meg, amely hárommillió személy Facebook-adatait tartalmazta, melyeket a MyPersonality appon keresztül gyűjtöttek be. Az adatkészlet tartalmazta az emberek életkorát, nemét, lokációját, állapotfrissítéseit, valamint egy személyiségjegyeket vizsgáló kérdőív eredményeit. (Az már csak hab a tortán, hogy egy banális hiba miatt lényegében a MyPersonalityvel gyűjtött összes adat kiszivároghatott, és emiatt a Facebook ki is tiltotta az appot.)

A kutatócsapat a modell alapján készített egy algoritmust is, amellyel a nyilvánosan elérhető adatok alapján és mindössze 15 attribútum felhasználásával az amerikai polgárok közel száz százalékát (99,89 százalék) képesek voltak beazonosítani.

Magyarán módszerükkel pontosan vissza lehetett fejteni például, hogy melyik amerikai állampolgár milyen válaszokat adott a népszámláláskor.

A kutatók létrehoztak egy oldalt is, ahol a módszerben kétkedők böngészőben kipróbálhatják egy korlátozott adathalmazon a szoftvert. Meg kell adni különböző adatokat (életkor, nem stb.), és a szoftver megmondja, hogy hány százalékos valószínűséggel azonosítható be az illető. A részletek csak az alapteszt kitöltése után válnak láthatóvá. (A Nature Communicationsben még az szerepel a cikk végén, hogy a kísérletek reprodukálásához szükséges forráskód is elérhető a dokumentációval, a tesztekkel és a példákkal együtt, de ennek már nincs nyoma az oldalon.)

Megoldás: majdhogynem nincs

A magánélet védelmének bevett módszere például az attribútumok eltávolítása vagy a hamis értékre cserélése, esetleg hogy egy adatkészletnek mindig csak egy töredékét teszik elérhetővé. A kutatók szerint azonban ezek sem elégséges módszerek a személyes adatok védelmére.

A másik véglet az ilyen adatkészletek teljes anonimizálása lenne, csakhogy akkor lényegében egy olyan – elemezhetetlen – adathalmot kapunk, aminek semmi értéke sincs a kutatók számára. Így például egy egészségügyi adatkészletnél lehetetlen lenne reprodukálni egy kutatócsapat eredményeit.

Megoldás jelenthet a hozzáférés szigorítása. Például érzékeny orvosi adatokhoz csak biztonságos és zárt körülmények között lehetne hozzáférni, ahol a másolásra sem lenne lehetőség. Erre már vannak kísérletek, például a franciák létrehoztak egy központot, amely interfészként kapcsolja össze az adatok előállítóit és felhasználóit. A CASD (Secure Data Access Centre) központ például azt ígéri, hogy ellenőrzött körülmények között és feltételekkel és csak célzottan lehet hozzáférni az adatokhoz. A központban nagyságrendileg 66 millió személy különböző adatait (az egészségügyitől a népszámlálási adatokig) tárolják. Ezekkel az adatokkal csak speciális hozzáférési pontokon lehet dolgozni.

Az elemzésnél megoldás lehet, hogy a nyers adatokat ún. multi-party titkosítással rejtik el (a kriptográfiai módszerről itt: https://bitport.hu/kis-magyar-ceg-indul-a-hsm-piac-meghoditasara írtunk). Ez elméletileg akár működhetne is, de például a tudományos kutatásnál ez sem feltétlenül járható. Mivel a kutató magukat a nyers adatokat nem látja, így fel sem ismerheti, ha hibázott valahol.

Forrás: https://bitport.hu/ez-az-algoritmus-feladja-a-lecket-a-gdpr-felelosoknek-senki-sem-maradhat-nevtelen

Válogatta: Fonyó Istvánné