„Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ

OCHRANA CITLIVÝCH ÚDAJOV V KONTEXTE DÁT ZO SIETE SMARTGRID

PREDPOKLADY A PODMIENKY, KTORÉ JE NUTNÉ NAPLNENIŤ, METÓDY, NÁSTROJE A AUTOMATIZÁCIA ANONYMIZAČNÝCH PROCESOV.

Predmetom štvrtej analýzy v rámci projektu je jeho bezpečnostné usmernenie v oblasti ochrany citlivých údajov. Táto správa je len formálnym výstupom analýzy vhodnosti rôznych prístupov k procesu anonymizovania a manažmentu osobných údajov v dátových štruktúrach potencionálneho systému sledovania pohybu výroby a spotreby elektrickej energie na Slovensku. Cieľom analýzy je popísať predpoklady a podmienky, ktoré treba splniť pri zbere, analýze, správe, uskladnení a distribúcii z inteligentných systémov estimácie odchýlky výroby a spotreby.

1. Všeobecne k ochrane osobných údajov v IS

Ochrana osobných údajov predstavuje jednu z hlavných priorít každého systému, ktorý zbiera, spracováva a uchováva informácie o svojich zložkách v takej forme, v ktorej sú tieto jeho zložky identifikovateľné externými entitami, ktoré k týmto informáciám nemajú oprávnený prístup.

Moderné systémy podporujúce súčasné bilančné modely estimácie odchýlky výroby a spotreby elektrickej energie v sieti nie sú výnimkou. Získanie dát a meta-dát o koncových užívateľoch, ktorými tieto systémy disponujú, predstavujú potenciálne veľký zdroj peňazí pre jednotlivcov a skupiny organizovaného kybernetického zločinu. Bohužiaľ, ochrana týchto údajov je vo svete často kriticky podhodnotená tak na úrovni legislatívnej (legislatívna definícia pojmu osobný údaj, aplikácia legislatívy) ako aj na úrovni operatívnej (práca s dátami, ich anonymizovanie, ich uskladnenie, fyzický prístup k dátam). Kybernetické útoky na vládne databázy európskych štátov zaznamenávajú v posledných rokoch prudký nárast. Dôvody tohto trendu sú tri: 1.) zvyšujúca sa počítačová gramotnosť populácie, 2.) znižujúca sa cena výpočtovej techniky a rast jej výkonu, 3.) uvoľňovanie štátnych dát v agregovanej a surovej podobe ako súčasť iniciatívy voľného prístupu k informáciám (OpenData), s nezvládnutým systémom anonymizovania týchto databáz. Viaceré štáty, na čele s Veľkou

Britániou, Ruskom, Nórskom a Nemeckom vyhlásili otvorenú diskusiu o ochrane osobných údajov a financujú internú agendu, ktorá má systém ochrany štátnych databáz modernizovať.

Na nasledujúcich stranách popíšeme túto problematiku najprv v jej teoretických základoch, v kontexte limitov anonymizovania, rôznych prístupov k riešeniu tohto problému a nakoniec v kontexte systému bilančného modelu estimácie odchýlky výroby a spotreby elektrickej energie po spustení inteligentnej siete (SmartGrid).

2. Osobné údaje

Pojem osobný údaj je slovenskou legislatívou identifikovaný prostredníctvom zákona 122/2013 Z. z. o ochrane osobných údajov. Jeho aktuálne znenie je kompatibilné so Smernicou 95/46/ES Európskeho parlamentu a rady EU o ochrane jednotlivcov, spracovania osobných údajov a o voľnom pohybe osobných údajov.

Definícia pojmu v plnom znení: Osobnými údajmi sú údaje týkajúce sa určenej alebo určiteľnej fyzickej osoby, pričom takou osobou je osoba, ktorú možno určiť priamo alebo nepriamo, najmä na základe všeobecne použiteľného identifikátora alebo na základe jednej či viacerých charakteristík alebo znakov, ktoré tvoria jej fyzickú, fyziologickú, psychickú, mentálnu, ekonomickú, kultúrnu alebo sociálnu identitu.

Definíciou údaju je jednotlivý fakt a osobný údaj je charakterizovaný ako súbor údajov, teda súbor jednotlivých faktov, na základe ktorých je osoba určiteľná alebo určená. Preto náhodný súbor údajov musíme charakterizovať ako osobné údaje vtedy keď je možné jednotlivca na základe týchto údajov identifikovať, či už priamo alebo nepriamo.

3. Náročná aplikácia legislatívnych definícií v praxi

Táto definícia osobných údajov je síce evidentná, no v aplikácii často nedostatočná. To, či je údaj osobný alebo nie do veľkej miery závisí od kontextu, v ktorom je uvedený. Pokiaľ je jednotlivec popísaný ako veriaci vo vzorke, v ktorej je väčšina veriacich, tento údaj môžeme len ťažko použiť na jeho identifikáciu. Pokiaľ by bol ale jeden z mála veriacich vo vzorke, stáva sa tento údaj dobrým nástrojom na jeho identifikáciu, a teda podľa legislatívnej definície jeho osobným údajom. Takýmto spôsobom sa môžu stať osobnými údajmi informácie, ktoré by sme za takéto pôvodne vôbec nepovažovali (napr. doba odchodu do práce, preferencia jedla, ľubovoľná aktivita, či pasivita v ľubovoľnom čase na ľubovoľnom mieste).

Treba si preto uvedomiť, že osobné údaje nie sú len štandardne používané údaje ako meno, priezvisko, rodné číslo, adresa, číslo poistenca či informácie o zdravotnom stave. Tieto informácie sú frekventovane identifikované ako osobné údaje len preto, lebo sa v súčasnom systéme často používajú ako identifikátory. V správnom kontexte môže byť identifikátorom ktorýkoľvek dostupný údaj o skupine ľudí. Problém je, že údaje ako číslo občianskeho preukazu alebo rodné číslo sami o sebe poskytujú aj pridané informácie o človeku, ktoré v danom kontexte nie sú potrebné, no uľahčujú identifikovanie konkrétnej osoby treťou stranou.

Riešením tohto problému by mohlo byť zavedenie nového štandardu v systémoch štátnej správy prostredníctvom MV SR, ktoré by do budúcna zabránilo takýmto kros-validáciám dát.

Z pohľadu štruktúry by malo ísť o alfanumerickú sadu znakov priradenú všetkým osobám evidovaným napríklad v registri obyvateľov. Takýto identifikátor nemôže obsahovať žiaden údaj popisujúci ľubovoľnú charakteristiku osoby, čím sa odlišuje od teraz používaných systémových identifikátorov ako rodné číslo, z ktorého možno odvodiť pohlavie, vek alebo rasu osoby. Takýto identifikátor by bol prirodzene menej náchylný na zneužitie kros-validáciu a vhodnejší ako kľúč pri procese anonymizovania dát.

4. Proces anonymizovania dát

Ak má ľubovoľný komunikačný systém vytvorený kvalitný proces anonymizovania dát, umožňuje mu to prenos dôležitých informácií cez miesta s nižšou informačnou bezpečnosťou, ako napríklad prenos medzi organizáciami. Anonymizovaný prenos bude disponovať relatívne nízkym rizikom zverejnenia citlivých informácií a stále umožňuje následnú analýzu dát a hodnotenie prenesených informácií.

Vo všeobecnosti, je anonymizovanie dát proces konvertovania čistých (surových) dát do sofistikovanej podoby, z ktorej surové dáta nemožno extrahovať reverzným procesom ich primárnej manipulácie. Podané inak, je to proces, v ktorom zoberieme dáta, z ktorých sme schopní identifikovať konkrétnu entitu (napr. osobu alebo skupinu osôb) a pretransformujeme ich na dáta, z ktorých sa tieto entity nedajú identifikovať.

K základným definíciám však treba doplniť niekoľko podmienok. Prvá, dáta anonymizované (za konkrétnym cieľom) podľa hore uvedených definícií môžu stále poslúžiť ako pomocný nástroj na identifikáciu entít, na ktorých ochranu sme sa pri anonymizovaní explicitne nezamerali. Druhá, na anonymizovanie dát sa používa súbor rozmanitých techník, ktorých cieľom je zachovať rozloženie dátového poľa (charakteristík dátového setu) ako jeho veľkosť, pozíciu dát, typ dát a ich frekvenciu. Dáta musia aj po anonymizovaní vyzerať realisticky, aby mohli byť použité v testovacom prostredí.

Z uvedeného vyplýva jeden veľmi dôležitý fakt. Z absolútne anonymizovaných dát (vynulovanie hodnôt) bude mať analytik len malý úžitok, nakoľko sa anonymizovaním znižuje ich hĺbka. Zmenšujú sa možnosti ich analýzy. Pri procese anonymizovania preto ide vždy o nájdenie kompromisu medzi bezpečnosťou a utilitou dát.

5. Základné metódy anonymizovania dát

Prvou, relatívne intuitívnou metódou anonymizovania je **vytvorenie náhradných symbolov**. Nahradenie číslic dátumu narodenia náhodnými znakmi je pre útočníka neprekonateľná prekážka. Je ale jasné, že takýto proces anonymizovania výrazne znižuje informačnú hodnotu dát. Anonymizovanie tvorbou náhradných dát podľa kľúča (nie náhodným priraďovaním symbolov) predstavuje skôr pseudo-anonymizovanie dát, nakoľko po obdržaní kľúča vieme dáta dostať do pôvodnej podoby. Existencia kľúča zvyšuje pravdepodobnosť úspešného útoku ale zároveň zvyšuje informačný potenciál dát.

Druhou metódou anonymizovania je **nulovanie dát.** Je to radikálny prístup k anonymizovaniu, pri ktorom sa citlivým dátam pripíše hodnota N/A alebo NULL. Týmto ťahom sa však informačná hodnota dát znižuje. Na druhú stranu, dáta takto anonymizované sú maximálne odolné voči nechcenej interpretácii. Tento prístup neodporúčame použiť na dáta, ktoré majú prejsť komlexnejšou post-anonymizačnou analýzou. Z takýchto dát vieme

totiž vyčítať maximálne ich frekvenciu výskytu v čase a ich objem. Podobne drastickým prístupom k anonymizovaniu, ktorý je dobré používať najmä pri extrahovaní za účelom odovzdania dát tretej strane nad ktorou nemáme kontrolu (nevieme na čo dáta potrebuje ani ako ich chce analyzovať) je odovzdanie dát v agregovanej forme. Zo surových dát sa spravia dáta s menšou granularitou, napríklad priemerovaním.

Treťou metódou anonymizovania je tvorba virtuálnych dátových vzoriek. **Virtuálne dátové vzorky** sú také vzorky, ktoré sú požadovanými, dopredu definovanými charakteristikami identické s pôvodnou vzorkou dát (napríklad poradím, typom, frekvenciou, rozložením dát), no v ostatných charakteristikách sú ich stavy nahradené inými hodnotami (napríklad vymyslené tel. čísla, dátumy narodenia). Zo štandardných metód anonymizovania dát je tvorba virtuálnych vzoriek asi tá najefektívnejšia. Na aplikáciu tohto typu anonymizovania potrebujeme vytvoriť knižnicu, z ktorej sa budú virtuálne údaje vyberať a pripisovať reálnym dátam a logické pravidlá, podľa ktorých ich budeme pripisovať jednotlivým stĺpcom. Anonymizovanie pomocou virtuálnych vzoriek poskytuje dobrý kompromis medzi zachovaním informačnej hodnoty dát po anonymizovaní a ich bezpečnosťou.

Klasickou chybou pri aplikácii metódy virtuálnych vzoriek je takzvaný data overflow. Stane sa to vtedy, keď pri tvorbe virtuálnej vzorky nedodržíme charakteristiky pôvodnej vzorky (zväčšíme-zmenšíme počet riadkov alebo stĺpcov alebo zmeníme distribúciu dát). Ak chceme virtualizovať naše dáta naozaj kvalitne, treba ich v prvom rade popísať so všetkými ich charakteristikami, ktoré potrebujeme zachovať. Napríklad rodné čísla sú všetky deliteľné číslom 11. Ak by sme ich chceli virtualizovať ale stále použiť ako identifikátor v systéme, ktorý kontroluje rodné čísla deliteľnosťou jedenástimi, musíme túto charakteristiku zachovať aj vo virtuálnych dátach. Z bezpečnostného hľadiska sa oplatí virtualizovať dáta tak, že sú na prvý pohľad na nerozoznanie od pôvodných dát. Číselné hodnoty nahradzujeme hodnotami v podobnej sústave, geografické názvy geografickými názvami a nie len náhodným zhlukom písmen.

Štvrtou metódou anonymizovania je **shuffling.** Je to prístup, pri ktorom prehádžeme originálnu vzorku dát po dopredu definovaných osách, podľa konkrétnych premenných (dátových reťazcov). Metóda vyžaduje prípravu a zamyslenie sa nad jej vhodnosťou. Aby bola efektívna, musí byť prehádzanie dát dostatočne komplikované. Podobné premenné by sa nemali prehadzovať ak je premenných takého typu v dátach málo (umožňuje to ľahké de-anonymizovanie). Metódu neodporúčame použiť pri vzorkách s malým počtom premenných. Pri dostatočne veľkých vzorkách, kde dáta predstavujú názvy, mená alebo iné nominálne premenné má anonymizovanie prostredníctvom shufflingu lepší výstup z pohľadu informačnej hodnoty dát ako anonymizovanie prostredníctvom virtuálnych vzoriek. Pri anonymizovaní numerických intervalových hodnôt a niektorých ordinálnych premenných vieme použiť **metódu posunutia**, pri ktorej posunieme hodnoty o vopred definovaný interval alebo o náhodný interval. Manipuláciou veľkosti intervalu vieme manipulovať zachovanie štatistickej hodnoty dát.

Vo všeobecnosti, predovšetkým ale pri shufflingu, by sa pri anonymizovaní dát mala venovať pozornosť hodnotám devijuúcim od priemeru, unikátne, nezodpovedajúce štandardnému rozloženiu dát vo vzorke. Platí tu pravidlo, čím sú dáta vo vzorke monotónnejšie, tým ľahšie je vzorka anonymizovateľná, čím viac štatistických výnimiek je vo vzorke dát, tým jednoduchšie je pre útočníka extrahovať zo vzorky parciálne dáta.

Pri implementácii základných anonymizačných techník je vhodné jednotlivé prístupy kombinovať. Správnym prístupom vieme za pomoci jednoduchých nástrojov anonymizovania dosiahnuť značnú úroveň bezpečnosti a zachovať pri tom interpretatívnu hodnotu dát v post-anaonymizačnej analýze. K návrhu a aplikácii systému procesov anonymizovania dát je treba pristupovať systematicky a s cieľom nájsť kompromis medzi ochranou osobných informácií a zachovaním analytického potenciálu týchto informácií.

Neriadené alebo prehnané anonymizovanie dát je často kontraproduktívne a nepotrebné lebo:

- Útočníci vo väčšine prípadov nemajú i tak všetky potrebné informácie na systematickú extrakciu dát.
- Nie každá informácia z databázy je citlivá v kontexte cieľu zberu týchto dát.
- Nie všetky dáta sú rovnako senzitívne v kontexte analýz a interpretácie, za ktorých účelom ich zbierame. Nie všetky vzťahy medzi premennými sú predmetom anonymizačného procesu. Napríklad vzťahy medzi používateľmi v on-line komunikačných systémoch sú často verejné, no vzťahy užívateľa k službe (napr. prihlásenie do služby, frekvencia používania) sú skoro vždy považované za senzitívne.
- Systém anonymizovania je vhodné od začiatku stavať na poznatkoch o tom, ktoré informácie v dátach sú a ktoré nie sú senzitívne. V projekte treba definovať čo presne je pokladané za osobný údaj a čo nie. Následne je nutné dopredu nadizajnovať plán pomeru anonymizovania dát proti strate ich informačnej hodnoty a riadiť sa ním počas procesu vývoja anonymizačného systému.

Na anonymizovanie dát sa používajú aj pokročilé prístupy a nástroje ako hashing, **syntaktické bezpečnostné modely** (K-anonymity, l-diversity, t-closeness, e-differential privacy) zameriavajúce sa na publikáciu dát a špecifikáciu ako by dáta mali vyzerať, ak majú byť zverejnené spolu s inými typmi dát, či systém **crowd blending privacy**. Každý z prístupov má samozrejme svoje výhody (väčšinou komplexitu) a nevýhody. Crowd blending privacy je napríklad ľahko dosiahnuteľný a dynamický, no neponúka silnú záruku bezpečnosti. Algoritmy typu **differential privacy** je náročné dizajnovať, ak majú po anonymizovaní na dátach bežať komplikovanejšie analytické úlohy, nakoľko potrebujú vlastnú optimalizáciu pre každú úlohu. Tento problém nedostatočnej utility majú aj syntaktické bezpečnostné modely (optimalizuj správne k pre k-anonymity, správne l pre l-diversity, správne e pre e-differential privacy pre každú úlohu). Problém s často používanou metódou **hashing** si popíšeme nižšie ako dobrý príklad procesu de-anonymizovania dát.

6. Pseudo-anonymizovanie dát

Ľudia si tradične zamieňajú anonymizované dáta za pseudo-anonymizované dáta. **Pseudo-anonymizácia** je proces v ktorom zoberieme surové dáta, v ktorých môžeme identifikovať osobné údaje, a konvertujeme ich do takého formátu, v ktorom tieto údaje nie je možné identifikovať. Zároveň disponujeme metódou, ktorou z týchto pseudo-anonymizovaných dát vieme konvertovať vzorku naspať do stavu, v ktorom vieme opäť identifikovať osobné údaje. Pseudo-anonymizovanie je teda reverzibilný proces, ktorý nám umožňuje uzamykať a odomykať hlbšiu informačnú a štatistickú hodnotu dát.

Tento prístup je dostatočný a dokonca často požadovaný, lebo v zamknutom stave disponujú dáta relatívne dobrou bezpečnosťou (riziko, ktoré je vždy prítomné je získanie kódovacieho kľúča útočníkom) a po odomknutí majú dáta svoj plný štatistický potenciál. Nebezpečenstvo predstavuje stav, keď si ľudia myslia, že ich dáta sú plne anonymizované a také ich ponúknu tretej strane, pri čom sú len pseudo-anonymizované. Naopak, plne anonymizovať dáta

v kontexte, v ktorom je ideálne či dostačujúce pseudo-anonymizovanie je zbytočný krok, ktorý okliešťuje ich analytický potenciál.

7. Proces de-anonymizovania a nedostatočné anonymizovanie dát

Napriek tomu, že podľa vyššie uvedených definícií, je proces pravého anonymizovania nereverzibilný, v reálnych aplikáciách existujú cesty, ako anonymizované dáta de-anonymizovať. Tieto metódy predstavujú hrozby pre systémy, v ktorých je anonymizovanie dát podmienkou. De-anonymizovanie je proces, ktorým z viacerých anonymizovaných databáz vieme pomocou **kros-referencie** a **kros-validizácie** extrahovať aspoň časť senzitívnych dát o užívateľoch systému. Problémom, vďaka ktorému je de-anonymizovanie možné, je vzrastajúci počet dát charakterizovaných ako osobné údaje (citlivé dáta) a zvyšujúci sa počet verejných zdrojov dát, ktoré môžeme podrobiť kros-referencii. Ako príklad, len na základe kros-referencie dát z verejných databázach USA, kde každá z nich je anonymizovaná pre svoj kontext, no ponecháva zostatkové (viditeľné) údaje, ktoré v tomto kontexte nie sú považované za citlivé údaje, je možné do hĺbky identifikovať 87% obyvateľov USA len na základe ich PSČ, veku a pohlavia. Legislatívne rozširovanie pojmu „osobný údaj“ preto nie je cestou k zvýšeniu bezpečnosti systému.

De-anonymizovaniu dát je možné sa brániť dodržaním niekoľkých podmienok:

i) uvedomme si a prijmime fakt, že riziko de-anonymizovania je reálne,

ii) tvorbu bezpečnostného systému vnímajme ako snahu o dynamické a kontinuálne minimalizovanie rizika,

iii) komunikovať tento prístup zúčastneným inštitúciám (často požadujúcim absolútne riešenia aj tam, kde nie sú možné alebo zmysluplné) je náročné, no čím viac zúčastnených to pochopí, tým ľahší bude vývoj a implementácia,

iv) racionálne limitujme detail dát,

v) počas vývoja, by sme sa mali opakovane pozrieť na dáta, ktoré plánujeme zo systému posielať tretím stranám a predstaviť si, ako by jednotlivci mohli byť týmito dátami identifikovaní.

Sú v dátach segmenty ktoré sú viac senzitívne ako sú potrebné pre spracovanie a interpretáciu dát? Ak áno, dáta sú nedostatočne chránené. Ak ste schopní agregovať jednotlivcov v dátach do stabilných kohort a malých skupín, ktoré reprezentujú priemerné správanie v dátach, dáta sú náchylnejšie k de-anonymizovaniu.

Ukážku nedostatočného anonymizovania dát, či neefektívne aplikovanej metódy popíšeme na frekventovane používanej metóde anonymizovania - metóde **hashing**. Pre pochopenie popis zjednodušíme. Predstavme si hashing ako funkciu, v ktorej na jednej strane dáme údaj (napr. rodné číslo) a na výstupe nám funkcia dá jedinečný identifikátor ako: k9089jthg876fe6534. Funkcia má tri dobré vlastnosti: a) pre ten istý vstup dodá stále rovnaký výstup, b) pre iný vstup dodá iný výstup, bez viditeľného vzťahu k vstupu prvému, c) hashing nie je reverzibilná funkcia, takže útočník ktorý by chcel získať rodné číslo z výstupu k9089jthg876fe6534 otočením procesu funkcie sa nikde nedostane. Tieto vlastnosti sú ušľachtilé, no dostatočne skúsený útočník vie stále hashing obísť. Stačí že má k dispozícii hashovaciu funkciu a vie, že boli hashované rodné čísla. Ďalej vie, že rodné čísla majú konkrétnu štruktúru. Stačí mu už len vygenerovať všetky existujúce kombinácie rodných čísel

(nie je ich až tak veľa), vložiť ich postupne do hashovacej funkcie a porovnávať výstupy s naším až kým neidentifikuje naše rodné číslo. Ešte pripravenejší útočník si vygeneruje kombinácie dopredu a vytvorí ich index, aby naše rodné číslo získal v priebehu niekoľkých sekúnd.

Príklad nedostatočného anonymizovania dát z energetiky predstavuje link http://open.enernoc.com/data/ kde sú k dispozícii 5-minútové dáta o spotrebe elektrickej energie od 100 anonymizovaných komerčných a industriálnych klientov. Samotné mená organizácií sú síce zakódované, no aspoň niektoré z nich bude možné identifikovať na základe metadát (sú tiež k dispozícii) o randomizovanej geografickej polohe, typu priemyselného odvetvia a rozlohy v stopách štvorcových. To je pochopiteľne neprijateľné.

8. Dizajn bezpečného systému

Pri dizajne a vývoji systému ktorý bezpečne pracuje s dátovým tokom sa musíme riadiť niekoľkými princípmi. Prvým je interoperabilita existujúceho a vyvíjaného softvéru. Aplikácie musia odpovedať na viacero typov dotazov, vzniká požiadavka, aby existujúci softvér a softvér vo vývoji bežal na anonymizovaných a pseudo-anonymizovaných dátach. Druhou požiadavkou bude pochopiteľne extenzibilita existujúceho softvéru, predpokladáme, že tento bude musieť generovať agregované štatistiky. Škálovateľnosť systémov je podmienkou, nakoľko anonymizované dáta budú mať aspoň takú veľkosť ako originálna vzorka dát, pričom počet agregátov a metadát požadovaných jednotlivými zložkami systému môže dramaticky narásť. Dôležitým princípom je kontinuálne vylepšovanie utility dát z dátového toku, nakoľko dáta nebudú používané v izolovanom systéme a dátové agregáty vyťahované pre potreby jednotlivých inštitúcií budú v závislosti na ich rozmanitosti potenciálnymi zdrojmi dát pre proces de-anonymizovania kros-validáciou. Služby poskytovania agregovaných dát zo systému by preto mali byť koordinované.

Pri tvorbe anonymizačného systému je potrebné otvoriť aktívnu diskusiu s vlastníkom dát, ktorý bude systém spracovávať a budúcim administrátorom tohto systému. Cieľom diskusie bude, aby všetky strany intuitívne chápali nepriamu úmeru vzťahu medzi mierou anonymizovania dát (bezpečnosťou) a použiteľnosťou dát v analýze. Skupina by mala spoločne definovať mieru tohto pomeru, ktorú sú ochotní prijať v kontexte modelov (napr. modelov estimácie odchýlky spotreby elektrickej energie na Slovensku) a na úrovni konkrétnych dátových parametrov. Výstupom diskusie by mala byť aj zadefinovaná metrika bezpečnosti dát a metrika ich utility.

Na druhej strane, systém má byť pripravený na jeho užívateľov a ich požiadavky. Užívateľ má byť informovaný, že dáta, ktoré dostane z databázy sú anonymizované alebo sú to agregované dáta s menšou granularitou. Obmedzenia a štandardný formát, v ktorom sa budú dáta poskytovať tretím stranám, musia byť kvalitne popísané, aby sa predišlo nedorozumeniam pri tvorbe nástrojov, ktoré budú tieto upravené dátové vzorky automaticky generovať a zasielať. V ideálnom prípade by mal byť vytvorený systém na mapovanie query a requestov na databázu, ktorý by pomohol v ich manažmente, ukazoval by zaťaženie databázy požiadavkami a umožnil by detekovať podozrivé požiadavky (napr. časté, neautorizované). S takýmto systémom by bolo možné vytvoriť interaktívny prístup k databázam, ktoré by poskytovali agregované dáta podľa požiadaviek užívateľov (ak by ich požiadavka spĺňala bezpečnostné podmienky), namiesto poskytovania objemných anonymizovaných dátových vzoriek, nad ktorými by systém po odovzdaní nemal kontrolu.

9. Podmienky ochrany citlivých údajov v inteligentnej elektrickej sieti

Jedným z cieľov Smart Grid iniciatívy je umožniť správcovským inštitúciám, užívateľom, dodávateľom a tretím stranám monitorovať a kontrolovať spotrebu elektrickej energie. Predpokladá sa, že vyššia frekvencia zbieraných dát spolu so silnejšími analytickými nástrojmi, interpretáciou a integráciou vnútro štruktúrnej komunikácie záverov analýz zefektívni systém pre vyššie menované strany. Umožní im robiť lepšie rozhodnutia, lepšie pochopiť potreby a požiadavky klientov a distribučných partnerov, čo bude mať pozitívny dopad na efektivitu distribúcie elektrickej energie.

Prirodzene, Smart Grid iniciatíva so sebou prináša sériu vážnych otázok spojených s ochranou osobných údajov koncových užívateľov, nakoľko dáta o spotrebe domácností budú zbierané so značnou frekvenciou a granularitou sekundárnych premenných (napr. geografická poloha smartmetru). Monitorovanie spotreby elektrickej energie je v princípe monitorovanie správania sa ľudí na dennej (minútovej) báze. Tieto dáta majú extrémnu hodnotu napríklad pre reklamné spoločnosti, štátne bezpečnostné zložky a kriminálne aktívne skupiny. Spreneverenie či de-anonymizovanie takéhoto typu informácie by zasiahlo do základných ľudských práv užívateľa a preto by mala byť tvorba kvalitného bezpečnostného systému na zber, manažment, analýzu a distribúciu dát o elektrickej výrobe a spotrebe absolútnym imperatívom iniciatívy Smart Grid.

10. Podmienky na vývoj bezpečnostného systému v kontexte Smart grid

V kontexte tvorby tohto projektu musíme zachovať stav, v ktorom bude možné pripísať dáta ku špecifickému smartmetru alebo OOM a zároveň ochrániť osobné údaje fyzických a právnických osôb, ktoré tieto dáta generujú svojim správaním. Tento cieľ musí byť splnený bez negatívneho dopadu na kontinuálny tok odmerných dát vo vysokej frekvencii a bez negatívneho dopadu na prebiehajúce operácie v sieti.

Z kontextu ďalej môžeme definovať tieto špecifikácie budúceho systému:

1. Odmerné dáta potrebné pre účtovanie spotreby a fakturácie musia byť pripísateľné OOM, bezpečne priradené ku konkrétnemu klientovi alebo majiteľovi účtu.

2. Pre účely účtovania je dostatočné generovať agregované štatistiky pripísateľných dát k OOM na dennej mesačnej a štvrťročnej báze.

3. Odmerné dáta pre účely estimácie odchýlky výroby a spotreby (pre bilančné systémy) a pre účely manažmentu siete nemusia byť pripísateľné konkrétnym domácnostiam, osobám či právnickým osobám. Anonymizované dáta sú dostatočné, pokiaľ: i) sú overiteľné ii) je ich možné priradiť ku konkrétnej správcovskej distribučnej stanici, spravujúcej sub-set smartmetrov (distribučné spoločnosti a sub-sety na úrovni okresov, mestských častí, či obytných blokov).

4. Anonymizované dáta budú zbierané s vysokou frekvenciou, aby bolo možné v reálnom čase reagovať na fluktuáciu dostupnej elektrickej energie v sieti a na potenciálne problémy s distribúciou v sieti.

5. Najmenšia (akceptovateľná) jednotka anonymizovaných dát o spotrebiteľoch pre účely estimácie odchýlky výroby a spotreby, je dátový agregát na úrovni distribučných staníc, z ktorých sa energia dodáva priamo jednotlivým koncovým konzumentom (napr. prvá úroveň nad smartmetrami domácností). Systém nepotrebuje vedieť aké dáta sa generujú ktorým smartmetrom, zaujímajú ho vlastnosti týchto dát a ich agregát v čase. Systém by však mal byť schopný spravovať jednotlivé smartmetre a ich funkcie (napr. vypnúť dodávku) aj keď užívateľ bude anonymný, nakoľko je to požiadavka zo strany manažmentu siete.

6. Podmienkou pre stavbu bezpečného systému je dôverný vzťah medzi podnikmi verejných služieb zodpovedných za správu siete, úložiskom dát a dodávateľmi smartmetrov. V súčasnosti nie je úplne jasné, ako bude na Slovensku postavené vlastníctvo dát generovaných smartmetrami v systéme Smart Grid a kto všetko bude mať oprávnenie pristúpiť k týmto dátam v čistej forme, či ich uchovávať. Kým podľa niektorých nariadení by mal klient (užívateľ) vlastniť dáta o vlastnej spotrebe a výrobe, nakoľko ich generuje vlastným správaním, iné už zabehnuté systémy dávajú formálne vlastníctvo dát na stranu inštitúcií spravujúcich sieť, či prekvapivo, poskytovateľom služieb. Prístup k dátam pre jednotlivé strany je väčšinou definovaný legislatívou na úrovni štátu, ktorá sa snaží prihliadať napríklad aj na prístup k dátam počas mimoriadnych situácií. Tieto vzťahy by mali byť jasné pred začatím vývoja bezpečnostného systému, ktorý, ako sme už uviedli, do značnej miery závisí od kontextu, do ktorého je zasadený. Na základe týchto špecifikácií je následne možné odvodiť pravidlá manažmentu a manipulácie dát v takomto systéme:

Zber dát by mal byť v každej časti systému prísne obmedzený len na dáta potrebné k takým operáciám, za ktoré každá inštitúcia v systéme Smart Grid zodpovedá (napr. plánovanie a manažment, testovanie efektivity distribúcie, výpočet bilancie, manažment účtov a fakturácia).

Uchovávanie dát, z ktorých je možné extrahovať osobné údaje či informácie o správaní užívateľov, by malo byť v každej časti systému prísne obmedzené na dobu potrebnú na splnenie aktivít s ktorými bol koncový užívateľ oboznámený a ktoré akceptoval (napr. aktivity potrebné na tvorbu predikcií budú zrejme predstavovať horný strop, lebo je na ne potrebné značné množstvo dát). Po tom, ako boli všetky schválené aktivity na dátach vykonané, inštitúcie sú povinné nenávratne zničiť pôvodné dáta, všetky ich kópie a metadáta z nich vytvorené.

Metodologickou podmienkou distribúcie dát je **zadefinovanie** a **zachovanie bezpečných dátových intervalov**. Aby sme mohli dobre implementovať princíp minimalizácie zbieraných dát, potrebujeme vedieť, čo vieme odvodiť zo zbieraných dát ako funkciu rôznych intervalov. Zistiť, ako funguje vzťah medzi veľkosťou intervalov zberu, množstvom informácií, ktoré z takýchto zberov vieme odvodiť a bezpečnosťou v kontexte Smart Grid siete, si vyžaduje ďalší výskum.

Okrem intuitívnej **agregácie dát v čase**, je vhodné vytvárať aj **agregáciu dát na základe užívateľov**. Bezpečnostné riziká spojené s meraním energetického spotrebného správania užívateľov vieme značne obmedziť, ak bude systém vytvárať agregáty na základe blízkosti alebo podobnosti užívateľov (napr. podľa obytných blokov, podľa podobného profilu spotreby etc.) a individuálne dáta anonymizovať. Aj tu je potrebné definovať do akej miery treba dáta agregovať a aký je v konkrétnych prípadoch správny pomer medzi bezpečnosťou a utilitou agregátov. Pritom treba mať na pamäti, že agregácia neuchráni jednotlivca, ak majú všetci v agregovanej vzorke podobné hodnoty vo viacerých premenných. Vtedy je metódu potrebné kombinovať s anonymizovaním iného typu. Minimálne charakteristiky vzorky, potrebné na to, aby bola agregácia efektívna sú: i) vzorka o veľkosti minimálne 15-ich užívateľov alebo ordinálne/nominálne rozlíšiteľných skupín užívateľov, ii) v každej rozlíšiteľnej skupine užívateľov nesmie byť jedna premenná, podľa ktorej je možné identifikovať len jedného užívateľa a zároveň ani jeden užívateľ nesmie generovať viac ako 15% všetkých dát, ktoré

budú spoločne agregované. Pokiaľ pravidlo 15/15 nie je v agregáte dodržané, takéto dáta neodporúčame posúvať tretím stranám, nad ktorými nemáme kontrolu.

Latencia dát. Na všetky dáta, ktoré budú odovzdávané priamo užívateľom, fyzickým osobám, či komerčným a nekomerčným právnickým osobám mimo systému, navrhujeme aplikovať princíp prístupu k dátam s latenciou. Jednou z obáv zneužitia dát o spotrebe elektrickej energie je, že budú použité na odhad neprítomnosti užívateľov v ich domácnostiach. Pokiaľ by systém tretím stranám ponúkal dáta s časovým odstupom a zároveň s upravenými časovými radami, tento problém by odpadol.

11. Analýza anonymizovania Smart Grid dát v zahraničných systémoch

Tvorba systému, ktorý zaručuje bezpečnú manipuláciu s dátami získanými v inteligentnej energetickej sieti je relatívne neprebádaná téma. Pri analýze už nasadených systémov sme zistili, že systémové regulácie a bezpečnostné požiadavky na dátový manažment sú často špecifikované príliš všeobecne, ak vôbec. Výsledkom je, že nastavenie implementačnej praxe pre systémy manipulácie dát je ponechané na verejnú správu nižšej úrovne, na komerčné a konzultačné firmy. Tieto inštitúcie si preto sami regulujú bezpečnosť dát s ktorými pracujú. Bohužiaľ, v niektorých prípadoch nezávisle na sebe a s nedostatočnou odbornosťou.

Vlády a organizácie, vydávajúce štandardy, vypracovávajú aj bezpečnostné štandardy a nariadenia, usmerňujúce vývoj systémov na ochranu osobných údajov v kontexte iniciatívy Smart Grid. Komplexné a profesionálne odporúčania v severoamerickom prostredí poskytli napríklad Výbor pre severoamerickú energetiku (North American Energy Standards Board), Národný inštitút technologických štandardov (National Institute of Standards and Technology) a Oddelenie energetiky spojených štátov amerických (U.S. Department of Energy) a The Smart Grid Policy Framework, vydaný Prezidentskou kanceláriou v roku 2011. Konštruktívnym je aj Zborník č. 2 expertnej skupiny Smart Grid pri Európskej komisii, popisujúci regulačné odporúčania pre bezpečnosť dát, manipuláciu dát a ochranu osobných údajov pre štáty Európskej únie.

Praktické inšpirácie pri plánovaní a vývoji bezpečnostných systémov pre Smart Grid odporúčame získavať z podobne orientovaných systémov v odvetviach manažmentu medicínskych dát a anonymizačných systémoch telekomunikačných spoločností. V prvom z menovaných odvetví je kvalitne vyriešený problém pomeru medzi bezpečnosťou dát a zachovaním ich štatistického potenciálu, v druhom odvetví existujú vhodné riešenia na bezpečnostný manažment konštantného toku dát zbieraných vo vysokej frekvencii. Telekomunikačné spoločnosti majú tiež kvalitné štandardizované postupy pre komunikáciu s tretími stranami. Konkrétne pre limitovanie informácií, ktoré zdroje osobných údajov o klientoch môžu sprostredkovať tretím stranám bez súhlasu spotrebiteľa (lebo pýtať sa spotrebiteľa o povolenie zakaždým, keď agregát s jeho dátami posúvame v systéme by bolo nerealizovateľné), pre rozhodnutia o tom, aké informácie o zákazníkoch môžu dať tretím stranám centrá zákazníckeho servisu a pravidlá oznamovacích povinností pre časti systému, ktoré disponujú dátami v takej forme, z ktorej ešte možno extrahovať osobné údaje užívateľov.

Bohužiaľ, aj v týchto oblastiach sa kontext, v ktorom bezpečnostné systémy pracujú, občas podcení. Dobrým príkladom je, ako telekomunikačné inštitúcie vnímajú telefónne číslo. Telefónne číslo samo o sebe nie je osobným údajom kým nie je verejne spojené s ďalšími

údajmi o jeho majiteľovi. Na Slovensku neexitovala služba, ktorá by pre telefónne číslo vrátila meno vlastníka linky, kým ju nezačal poskytovať Slovak Telekom. Týmto krokom spoločnosť premenila telefónne čísla, pôvodne bežne dostupný údaj o svojich klientoch, na osobný údaj, ktorý je však stále bežne dostupný.

12. Zhodnotenie

Našim cieľom v tejto časti analýzy bolo urobiť prehľad problematiky bezpečnosti zberu, manažmentu, analýzy uchovávania a distribúcie dát získaných v pripravovanom systéme Smart Grid. Následne testovať rôzne prístupy k anonymizovaniu týchto dát na sprostredkovanej vzorke. Technický popis tohto procesu je príliš komplexný pre potreby tohoto reportu, ktorý zaraďujeme ako výstup projektu, preto report obsahuje hlavné závery, odporúčania a špecifikácie, ktoré sme extrahovali počas našej práce.

Anonymizačné metódy sú veľmi užitočné nástroje, ale ako je to so všetkým v paradigme bezpečnosti a správe osobných údajov, kľúčom k tvorbe odolného systému je pochopenie kontextu, v ktorom bezpečnostný systém bude aplikovaný. Zložky aktívne integrované do systému musia plne porozumieť svojmu rozhodnutiu anonymizovať svoje dáta a aký má a nemá daný proces anonymizovania vzťah k ochrane osobných údajov užívateľov.

V procese prípravy realizácie projektu bude nutné identifikovať, ktoré údaje zo Smart Gridu potrebujú jednotlivé inštitúcie, za akým cieľom a komu tieto údaje budú sprostredkovávať. Po tejto fáze bude možné identifikovať pomer medzi mierou anonymizovania a mierou ponechania štatistického potenciálu dát pre každú z inštitúcií a pre ich vzájomnú dátovú interakciu. V záujme vytvorenia kvalitného bezpečnostného systému je nutné si vyjasniť, aké sú potreby jednotlivých inštitúcií v systéme.

Na strane vývoja IS je nutné prepojiť teoretické poznatky o anonymizovaní dát s praxou v energetike. Posledný cieľ je síce náročný a dlhodobý, no nemôžeme ho nespomenúť. Z analýzy podobných bezpečnostných systémov sa ukazuje, že jedno z najslabších miest systému je samotný užívateľ, ktorý si vypýta dáta o svojej spotrebe (na čo má nárok) a následne ich sprostredkuje tretej strane, ktorá má potenciál ich zneužiť. Vzdelávanie koncových užívateľov v oblasti bezpečného narábania s dátami o svojej spotrebe preto má potenciál eliminovať značné množstvo útokov.

13. Záverečné poznámky

V rámci zhodnotenia vyššie napísaného uvádzame ešte nasledovné poznámky:

- Zložky systému by mali mať prístup k údajom o užívateľoch v takých agregátoch, ktoré im umožňujú realizovať procesy, ktorými sú poverené (napr. optimalizácia a manažment siete, koordinácia s distribučným a prenosovým systémom, účtovné a fakturačné služby, tvorba predikcií do bilančného systému).

- Klienti by mali byť schopní sa rozhodnúť, či poskytnú svoje dáta o spotrebe tretím stranám za účelom iným ako sprostredkovanie elektrickej energie.

- Všetky zložky systému sú schopné pracovať s kvalitne extrahovanými agregátmi dát (vytvorenými špecificky pre ich účel) a neexistuje dôvod, prečo by mali mať

k dispozícii surové dáta. Keď bude ktorákoľvek zložka pracovať so surovými dátami, bezpečnosť celého systému sa značne zníži.

- Zložky systému by nemali dávať agregované dáta k dispozícii tretím stranám bez súhlasu užívateľov.

- Autorizačný systém by mal byť postavený na princípe autorizačného procesu, v ktorom jednotlivé zložky systému a tretie strany dostavajú autorizáciu na prístup k dátam na limitovaný čas, a iba na základe ich požiadavky, v ktorej musí byť jasne uvedený účel, pre ktorý o dáta žiadajú.

- Pokiaľ je tretia strana autorizovaná na prístup k dátam, jej povinnosťou je tieto dáta chrániť, nedistribuovať a iniciovať na nich len operácie, ktoré boli popísané v žiadosti o autorizáciu. Nedodržanie týchto limitov by malo byť penalizované.

14. Otvorené otázky

Z uvedeného vyplývajú nasledovné otázky, ktoré v súčasnosti, ešte nie sú zodpovedané:

- Kto bude vlastníkom dát generovaných smartmetrami v sieti Smart Grid na Slovensku?

- Kto bude zodpovedný za správu systému ochrany osobných údajov a bezpečnostného manažmentu dát (napr. tvorby agregátov pre jednotlivé zložky systému, užívateľov a tretie strany)?

- Budú mať klienti možnosť odmietnuť inštaláciu smartmetru alebo moderovať, aké dáta v akom intervale sú o nich zbierané (v Kanade a niektorých štátoch U.S.A. takú možnosť majú)?

- Aké možnosti budú mať užívatelia, ak chcú ohlásiť problém so smartmetrom alebo únik informácií o ich osobe?

- Akú rolu budú v systéme hrať štátne bezpečnostné zložky Ministerstva vnútra SR a Ministerstva obrany SR (budú priamou súčasťou systému alebo budú figurovať ako tretia strana ktorá pre prístup k dátam potrebuje autorizáciu)?

- Aké dáta budú systémové zložky schopné odovzdať autorizovaným tretím stranám? Akú kontrolu bude mať koncový užívateľ nad odovzdávaním dát o jeho správaní zložkami systému tretím stranám?

- Za akých podmienok môže legislatíva zakázať poskytovať dáta tretím stranám či zložkám systému?

- Aké informácie bude nutné predložiť spolu so žiadosťou o autorizáciu prístupu k dátam?

- Akými nástrojmi budú užívatelia vzdelávaní o bezpečnej manipulácii s ich dátami?