VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ BAKALAURO STUDIJŲ PROGRAMA

BDAR įgalinantys sprendimai duomenų modeliuose: duomenų anonimizacija

Solutions to Enable GDPR in Data Models: Data Anonymization

Bakalauro baigiamasis darbas

Atliko: Valentinas Ešvovičius (parašas)
Darbo vadovas: Dr. Agnė Brilingaitė (parašas)
Darbo recenzentas: Dr. Asta Vaitkevičienė (parašas)

Vilnius – 2020

Santrauka

BDAR, įsigaliojęs 2018 metais, kelia griežtus reikalavimus duomenų saugojimui, bet vis dar trūksta sprendimų įgalinančių BDAR ir taikančių duomenų anonimizaciją reliacinėse duomenų bazių valdymo sistemose. Šio darbo metu tirti egzistuojantys duomenų anonimizacijos sprendimai, sukurtas ir ištestuotas duomenų anonimizacijos prototipas, taikantis apibendrinimo, iškraipymo, slopinimo ir sukeitimo technikas, skirtas *PostgreSQL*. Tirtas anonimizacijos poveikis duomenimis, pradinių ir anonimizuotų duomenų palyginimo rezultatai rodo, kad skaitinių reikšmių pakeitimas į rėžius ar netvarkos pridėjimas skaičiams nepadaro didelės įtakos duomenų panaudojimui. Remiantis tyrimo rezultatais, anonimizuojant duomenų modelį svarbu pasirinkti tinkamus anonimizacijos metodus, darbe suformuluotos rekomendacijos jų pasirinkimui ir taikymui.

Raktiniai žodžiai: bendrasis duomenų apsaugos reglamentas, BDAR, duomenų anonimizacija, duomenų anonimizacijos technikos.

Summary

GDPR, which came into force in 2018, imposes strict data privacy requirements, but there is still a lack of solutions that enable GDPR and apply data anonymization in relational database management systems. In this work, the existing data anonymization solutions were investigated, and a data anonymization prototype using generalization, distortion, suppression, and swapping techniques for *PostgreSQL* was developed and tested. The effect of anonymization on the data studied, and the results of the comparison of the initial and anonymized data show that changing the numerical values to range or adding a noise to the numbers do not significantly affect the use of the data. Based on the results of the research, when anonymizing the data model, it is important to choose the appropriate anonymization methods, and the work formulates recommendations for their selection and application.

Keywords: General Data Protection Reglament, GDPR, data anonimization, generalization, swaping.

TURINYS

Įvadas

Šiuolaikiniame versle skaitmeninių duomenų kiekiai yra dideli ir vertingi, bet neapdoroti duomenys negali būti panaudojami [PAP18]. Vis tik ne visi yra susirūpinę savo duomenų saugumu. S. Greengard [Gre18] tyrė po 1000 Vokietijos, Didžiosios Britanijos ir JAV gyventojų. Visose trijose šalyse žmonės sutiktų parduoti savo duomenis už 130–140 eurų atlygį per mėnesį, o duomenų pardavimas keltų grėsmę žmonių privatumui.

2018 metų gegužės 25 dieną įsigaliojęs BDAR (Bendrasis duomenų apsaugos reglamentas) pakeitė nuo 1995 metų galiojusią duomenų apsaugos direktyvą. 2016 metais Europos Sąjungos atstovams paskelbus apie naująjį duomenų saugojimo reglamentą Europos ir viso pasaulio verslininkai, piliečiai ir reguliatoriai pradėjo ruoštis naujojo įstatymo įsigaliojimui. Didžiųjų duomenų kiekio augimas yra viena iš didžiausių problemų, su kuriomis susiduriama šiame amžiuje, o didžiųjų duomenų analizė suteikia tiek naudos, tiek žalos [Zar16]. Didėjantys duomenų kiekiai didina susidomėjimą duomenų analize dėl sukuriamų naujų verslo galimybių [MBR+19]. Pavyzdžiui, iš elektroninės prekybos svetainių surinktų duomenų galima profiliuoti klientus pagal jų ankstesnes paieškas ir pirkimus, o pagal gimimo datą, pašto kodą ir lytį galima atpažinti beveik 60% asmenų. BDAR tikslas – apsaugoti asmeninius žmonių duomenis ir sumažinti galimybes identifikuoti asmenis.

Naujasis duomenų apsaugos įstatymas ne tik pakeitė duomenų saugojimo taisykles, bet ir numatė jų nesilaikymo pasekmes: dideles pinigines baudas, galinčias siekti net 4% praėjusių metų įmonės pasaulinių pajamų arba nuo 10 iki 20 milijonų eurų [18]. BDAR stiprina ne tik anksčiau galiojusios duomenų apsaugos direktyvos taisykles, o ir sukuria naujas, kaip pavyzdžiui, galimybę būti užmirštam. Senoji direktyva buvo palikusi skirtingų interpretavimo galimybių perkeliant šį teisės aktą į kiekvieną šalį. Kadangi BDAR yra ne direktyva, o reglamentas, jis įgalina tokias pačias duomenų apsaugos taisykles visose Europos Sąjungos šalyse.

Mokslinio tyrimo, kurio metu buvo apklausti 141 aukštųjų mokyklų studentai, rezultatai rodo, kad žmones reikia supažindinti su BDAR taisyklėmis [MDK19]. Tyrimo rezultatai parodė, kad studentai nežino, kur bus naudojami jų duomenys, kaip ilgai jie bus saugojami ar kas gali pasiekti šiuos duomenis. Visuomenė nuolat turi būti informuojama apie savo asmeninių duomenų svarbą, o BDAR ne tik sukuria taisykles duomenų saugojimui, bet ir skatina žmones rūpintis savo duomenų saugumu.

P. Breirbath [Bre19] išskyrė BDAR poveikį praėjus metams po jo įsigaliojimo – 2019 metų vasario duomenimis duomenų apsaugos pareigūnai sulaukė 206326 bylų. Beveik pusė jų buvo susiję su skundais, o daugiau nei ketvirtadalis – su duomenų nutekėjimais. Dažniausiai skundai

buvo susiję su telemarketingu, reklaminiais el. laiškais ir uždaros grandinės televizijomis (angl. closed-circuit television). Taip pat, 2019 metų birželį 73% apklaustųjų europiečių buvo girdėję apie BDAR. Daugiausiai girdėję apie BDAR europiečiai gyvena Švedijoje, Olandijoje ir Lenkijoje, o mažiausiai – Prancūzijoje, Italijoje ir Belgijoje. Be to, BDAR turi įtakos panašiems įstatymams visame pasaulyje: Brazilijoje pirmasis šalies bendrasis duomenų apsaugos įstatymas pradės galioti 2020 metų rugpjūtį. Kalifornijos valstijoje JAV Kalifornijos vartotojų privatumo aktas įsigaliojo nuo 2020 metų sausio 1 dienos, o 17 kitų valstijų ruošia panašius įstatymus. Norėdamos tęsti savo verslus Europoje kai kurios šalys Afrikoje ir Pietryčių Azijoje taip pat sustiprino duomenų apsaugos įstatymus. Pietų Korėja naujina savo šalies teisės aktus, kad jie taptų panašūs į BDAR, o Indijos parlamentas svarsto naujas duomenų apsaugos taisykles. Taigi, BDAR ne tik pakeitė duomenų apsaugos taisykles Europoje, bet ir skatina kitas valstybes keisti savo duomenų apsaugos įstatymus.

Šiame darbe nagrinėjama **problema** – BDAR įsigaliojo prieš beveik du metus ir kelia griežtus reikalavimus duomenų saugojimui, bet vis dar trūksta sprendimų įgalinančių BDAR ir taikančių duomenų anonimizaciją reliacinėse duomenų bazių valdymo sistemose. **Darbo tikslas** – sukurti ir ištestuoti asmeninių duomenų anonimizacijos prototipą reliacinėje duomenų bazių valdymo sistemoje.

Šio darbo **uždaviniai** yra:

1. Išskirti duomenų modeliavimui aktualias BDAR taisykles, jų įtaką verslui ir technologijoms.
2. Apžvelgti egzistuojančius duomenų anonimizacijos sprendimus.
3. Pasirinkti duomenų modelį ir užpildyti jį realiais arba generuotais duomenimis.
4. Įgyvendinti duomenų anonimizacijos metodų rinkinį ir pritaikyti jį pasirinktam duomenų modeliui.
5. Ištirti anonimizacijos poveikį pasirinkto modelio duomenims ir jų panaudojimui.

1. BDAR taisyklės ir asmeniniai duomenys

Asmeniniai duomenys turi ne vieną apibrėžimą. BDAR dokumente [18] asmeniniai duomenys yra „bet kokia informacija apie fizinį asmenį, kurio tapatybė nustatyta arba kurio tapatybę galima nustatyti (duomenų subjektas); fizinis asmuo, kurio tapatybę galima nustatyti, yra asmuo, kurio tapatybę tiesiogiai arba netiesiogiai galima nustatyti, visų pirma, pagal identifikatorių, kaip antai vardą ir pavardę, asmens identifikavimo numerį, buvimo vietos duomenis ir interneto identifikatorių arba pagal vieną ar kelis to fizinio asmens fizinės, fiziologinės, genetinės, psichinės, ekonominės, kultūrinės ar socialinės tapatybės požymius“. Asmeniniai duomenys taip pat yra vadinami ir asmenine informacija, kuri nurodoma ir kaip asmenį identifikuojantį informacija (AII) [VH19]. Vieni autoriai AII įvardija kaip bet kokią informaciją, kuri yra susijusi su konkrečiu žmogumi arba namų ūkiu. Kiti kaip rinkinį susidedantį iš vardo (arba pirmosios vardo raidės), pavardės ir bent vieno iš šių duomenų elemento: socialinio draudimo numerio, vairuotojo pažymėjimo numerio, kreditinės kortelės numerio, sveikatos apsaugos identifikavimo numerio, elektroninio pašto ir jo slaptažodžio. Taigi, nors įvairūs šaltiniai skirtingai įvardina, kas yra asmeniniai duomenys, juos sieja tai, kad asmeniniai duomenys yra tokie duomenys, iš kurių galima atskleisti žmogaus tapatybę.

Įsigaliojęs BDAR pakeitė nusistovėjusias duomenų saugojimo praktikas. Pagrindiniai BDAR duomenų apsaugos principai skelbiami II skyriaus 5 straipsnyje [18]:

- Teisėtumo, sąžiningumo ir skaidrumo: duomenų subjekto atžvilgiu duomenys turi būti tvarkomai sąžiningai, teisėtai ir skaidriai.
- Tikslo apribojimo: duomenys turi būti renkami aiškiais ir teisėtais tikslais.
- Duomenų kiekio mažinimo: asmeniniai duomenys turi būti renkami tik tokie, kurių reikia norint siekti tikslų, dėl kurių jie yra tvarkomi.
- Tikslumo: netikslūs duomenys turi būti atnaujinami arba ištrinami.
- Saugojimo trukmės apribojimo: duomenys turi būti ištrinami, jeigu jie yra nebenaudojami.
- Vientisumo ir konfidencialumo: turi būti užtikrinamas duomenų saugumas.
- Atskaitomybės: duomenų valdytojas yra atsakingas, kad būtų laikomasi aukščiau vardytų principų.

Taip pat, BDAR sukuria ne tik duomenų saugojimo principus, bet ir suteikia vartotojams teises bei sukuria pareigas įmonėms, kurios tvarko duomenis:

- Norint tvarkyti duomenis, turi būti įrodoma, kad vartotojas davė tam sutikimą „laisva valia“. Be to, atšaukti savo sutikimą turi būti taip pat lengva kaip ir duoti jį.
- Duomenų subjektas gali prašyti gauti duomenis, susijusius su juo, bei perkelti juos iš vieno

duomenų valdytojo pas kitą.

- Vartotojas gali nepagrįstai prašyti duomenų valdytojo ištrinti jo duomenis, o duomenų valdytojas yra įsipareigojęs ištrinti duomenis, jeigu jie yra nebereikalingi ar tvarkomi neteisėtai.
- Vartotojas gali prašyti apriboti duomenų tvarkymą, jeigu duomenys yra netikslūs, duomenų tvarkymas yra neteisėtas, duomenys yra nebereikalingi.

Taigi, BDAR ne tik kuria taisykles duomenų saugojimui, bet ir suteikia vartotojams galimybes labiau valdyti asmeninius duomenis.

2. BDAR įtaka verslui ir technologijoms

Įsigaliojęs BDAR turėjo daug įtakos. Pasak teisės žinovų, didžiausia naudą, kurią atneša BDAR Europos Sąjungos piliečių asmeniniams duomenimis yra, kad BDAR yra ne direktyva, o reglamentas [PAP18].

BDAR išplečia duomenų apsaugos apimtį taip, kad bet kuris žmogus ar organizacija, kuri renka ir apdoroja ES piliečių duomenis, turi vadovautis BDAR nepaisant to, kur yra saugomi duomenys [Tan16]. BDAR sukuria ir taisykles duomenų nutekėjimo atveju. Organizacijos, kurių duomenys yra nutekinami, turi pranešti duomenų apsaugos įmonėms per 72 valandas nuo nutekėjimo suradimo. BDAR taisyklių tikslas yra suteikti ES piliečiams supratimą, kas naudoja jų duomenis ir galimybę juos kontroliuoti patiems [Hsu18]. Organizacijos turi būti pasiruošusios lengvai ištrinti, tikrinti ir naudoti duomenis. Taip pat, joms rekomenduojama nerinkti duomenų, kurių įmonei nereikia ir ištrinti jau esančius nenaudojamus duomenis. Renkant visus duomenis vartotojams reikia paaiškinti, kam tie duomenys reikalingi, o surinktų duomenų negalima naudoti kitais tikslais, nei jie buvo rinkti. Pavyzdžiui, jei įmonė paprašo vartotojui suteikti jo telefono numerį su tikslu naudoti jį dviejų lygių autorizavimo sistemoje, jie jo negali naudoti kitiems tikslams.

Tyrimo kompanijos *Ovum* atliktoje apklausoje 2016 metais 52% organizacijų manė, kad BDAR lems baudas, o 68%, kad reglamentas stipriai didins verslo kainą Europoje [Tan16]. Dalis apklaustų įmonių teigė, kad BDAR įvedimas didins jų kaštus 10%. Tankard teigia, kad, apklausos duomenimis, šifravimas bei analizės ir ataskaitų teikimo technologijos yra laikomos geriausiomis technologijomis, norint pasiekti BDAR keliamų taisyklių įgyvendinimą. Kitos apklausos metu buvo vertinami svarbiausi įrankiai norint šifruoti duomenis. Daugiau nei 60% apklaustųjų išskyrė duomenų nutekėjimo prevenciją, daugiau nei 50% minėjo reikalavimų ir audito įgaliojimų įgyvendinimą bei finansinio bei kito turto apsaugą. Šifravimas turėtų apimti ne tik duomenų bazėse, bet ir skaičiuoklėse (angl. spreadsheets), tekstiniuose dokumentuose, prezentacijose, el. laiškuose ir archyvuose esančius duomenis. Nors šifravimas yra geras duomenų apsaugos įrankis, įmonės turi stengtis mažinti kaupiamų duomenų ir žmonių, turinčių prieigą prie duomenų, kiekį.

BDAR sukuria ne vieną iššūkį duomenų saugojimui. Asmeniniai asistentai (angl. personal assistants): *Siri*, *Cortana*, *Alexa*, turėjo didinti programų sudėtingumą dėl asmenį identifikuojančios informacijos išdavimo [Gre18]. Pokalbių botų (angl. chat bots), robotai patarėjai, rekomendacijų servisai susiduria su kitais informacijos saugojimo iššūkiais. Visos šios sistemos saugo duomenis, o ankščiau joms nereikėjo nustatyti, kur žmogus gyvena. Įsigaliojus BDAR gyvenamoji vieta tapo kertine informacija. S. Greengard darbe teigiama, kad BDAR gali stabdyti inovacijų plėtrą. Su šiais iššūkiais susidurs autonominės transporto priemonės, robotai ir kitos sistemos, kurios kliau-

nasi dirbtiniu intelektu. Įmonėms gali tekti laikyti duomenis atskirose duomenų bazėse, iš kurių viena būtų skirta ES piliečių duomenims, o kita – viso likusio pasaulio, arba rasti būdą kaip atskirti duomenis vienoje duomenų bazėje.

Vienas iš labiausiai matomų pakeitimų vartotojams Europoje įvedus BDAR buvo išaugęs prašymų kiekis tinklalapiuose sutikti su privatumo politika. Degeling ir kt. [DUL+18] atliko tyrimą, kurio metu lygino sutikimo su privatumo politikos pranešimų atsidarius tinklalapį kiekį ir jo pokytį populiariausiose puslapiuose visose 28 ES narėse. Tyrimui buvo atrinkti 500 populiariausių puslapių visose 28 šalyse. Pirmasis įvertinimas įvyko 2017 metų gruodį, nuo 2018 sausio iki balandžio tinklalapiai buvo vertinami kartą per mėnesį, o gegužę – 3 kartus (2 kartus prieš įsigaliojant BDAR ir kartą po įsigaliojimo), o tyrimo metu buvo atliekami ne tik automatiniai, bet ir rankiniai įvertinimai. Kadangi kai kurie tinklalapiai buvo populiariausiųjų sąraše ne vienoje šalyje, iš viso buvo tikrinami 6759 tinklalapiai. Lyginant įvertinimų duomenis sausį ir gegužę, jau įsigaliojus BDAR, buvo matomas 4.9% prieaugis, dauguma (79.6%) puslapių turėjo pranešimą apie privatumo politiką jau 2018 sausį, didžiausi prieaugį rodė šalys, kuriose sausį buvo fiksuoti prastesni rodikliai. Taip pat, lyginant privatumo politikų turinį 2017 kovo ir 2018 gegužės mėnesiais, buvo nustatyta, kad 85.1% privatumo politikų buvo keistos bent kartą, o jose augo ir sąvokų, susijusių su vartotojų teisėmis, kiekis. Be to, tyrimo rezultatai rodo, kad BDAR turi poveikį ne tik Europos, bet ir kitoms šalimis. Pavyzdžiui, 53% JAV ir 48% Rusijos 500 populiariausių tinklalapių pateko bent į vienos ES šalies 500 populiariausių tinklalapių sąrašą. Taip pat, prašymuose sutikti su privatumo politika buvo nustatytas padidėjęs kiekis terminų, kurie yra susiję su BDAR.

Įmonės norėdamos atitikti BDAR taisykles ir reikalavimus turėjo išleisti dideles pinigų sumas [APB18]. Visgi, baudos už BDAR nesilaikymą būtų buvusios dar didesnės. Nors išleista pinigų suma priklausė nuo daug įvairiu faktų, didžiausią įtaką darė įmonių dydis ir darbuotojų skaičius. Tyrime teigiama, kad įmonės, kuriose dirba nuo 1000 iki 5000 darbuotojų vidutiniškai išleido 1 milijoną Didžiosios Britanijos svarų sterlingų (GPB), įmonės su 50000 – 100000 darbuotojų – vidutiniškai 8 milijonus GPB, o įmonės su daugiau nei 100 tūkstančių darbuotojų – 49 milijonus GPB. Vidutiniškai investicijos vienam įmonės darbuotojui siekė 300 – 450 GPB. Didžiausias išlaidas įgyvendinant BDAR reikalavimus patyrė bankai (vidutiniškai 66 milijonai GPB), prekių, energetikos ir komunalinių paslaugų įmonės (vid. 19 mln. GBP) ir mažmeninių prekių prekybos įmonės (vid. 15 mln. GBP). Didelį BDAR poveikį pajautė dvi didžiausios pasaulio ekonomikos galios: Kinija ir JAV, nes abi šalys turi daug įmonių, kurios veikia ir ES šalyse [LYH19]. 2017 metais prognozuota, kad 68% JAV įmonių turėjo išleisti tarp 1 ir 10 milijonų JAV dolerių, o 9% įmonių – daugiau nei 10 mln. JAV dolerių. Šios išlaidos iš karto turėjo didinti paslaugų kainas klientams. Be to, ne visos įmonės nusprendė laikytis BDAR: išmanaus apšvietimo Kinijos įmonė *Yeelight* nustojo tiekti

savo paslaugas ES šalyse. O tokios įmonės kaip *Facebook* ir *Google* buvo paduotos į teismą kelios valandos po BDAR įsigaliojimo. Taigi, BDAR įsigaliojimas įmonėms reiškė didesnes finansines išlaidas.

BDAR įsigaliojimas turi įtakos ir kibernetinei saugai [LYH19]. Kadangi per pastaruosius metus įvyko daug duomenų nutekėjimo įvykių, o įsigaliojus BDAR apie pastebėtą duomenų nutekėjimą reikia pranešti per 72 valandas, įmonės turi stiprinti savo kibernetinį saugumą, o tai didins kibernetinio saugumo ekspertų poreikį. Tai reiškia, kad įmonės ir vyriausybės turi daugiau investuoti į kibernetinio saugumo švietimą. Pastarųjų metų pranešimai apie asmeninių duomenų saugumo pažeidžiamumą ir įvykiai, kai įmonės netinkamai naudoja ir pardavinėja vartotojų duomenis, mažino žmonių pasitikėjimą, o tai yra viena iš didžiausių šiuolaikinio verslo problemų. Tyrime teigiama, kad 39% vartotojų išleidžia daugiau pinigų, jeigu yra užtikrinti savo duomenų saugumu internete. Taigi, įgytas vartotojų pasitikėjimas įmonėms gali reikšti didesnes pajamas, bet, norint jį įgyti, reikia gerinti kibernetinį saugumą ir BDAR skatina tai daryti.

Įsigaliojęs BDAR turi įtakos ir internetinių tinklalapių srautui. S. Goldberg ir kt. [GJS19] tyrė BDAR įtaką Europos tinklalapių srautui ir elektroninės komercijos pardavimams. Jie naudojo žiniatinklio analizės duomenis iš 1508 įvairių įmonių, naudojančių *Adobe Analytics* platformą, rinkinio. Straipsnio autoriai pastebi, kad įsigaliojus BDAR, apdoroti žiniatinklio analizės duomenis galima tik gavus tam sutikimą. Tyrimo rezultatai parodė, kad įsigaliojus BDAR 4% sumažėjo puslapių peržiūrų per savaitę skaičius, o pajamos per savaitę mažėjo 8%. Taigi, BDAR įsigaliojimas kai kuriems puslapiams reiškė sumažėjusias pajamas.

BDAR daro įtaką ir vis populiarėjančioms šių dienų technologijoms: dirbtiniam intelektui ir blokų grandinėms [LYH19]. Li ir kt. teigia, kad BDAR apribos dirbtinio intelekto projektų apimtį ir didins jų kaštus. Kadangi vartotojai įsigaliojus BDAR turi galimybę ištrinti savo duomenis, tai gali turėti tiesioginę įtaką dirbtinio intelekto projektams, nes tai mažintų jų algoritmų efektyvumą ir tikslumą. Blokų grandinėse kiekviename bloke esanti informacija daro įtaką vėlesniems blokams. Jeigu vartotojai galės ištrinti duomenis ar juos pakeisti, tai gali mažinti blokų grandinių technologijos naudą ir efektyvumą. Seo ir kt. tyrė, kaip BDAR paveikė daiktų internetą. Kaštai didėja duomenų nutekėjimo atveju [SKP+18]. Jie teigia, kad vienas duomenų nutekinimo atvejis vidutiniškai įmonei kainuoja 150 JAV dolerių vienam gyventojui. Autoriai teigia, kad įmonių išlaidos, įsigaliojus BDAR, gali augti nuo 3–4 kartų iki 18. Taigi, BDAR kelia iššūkius ir populiariausioms šių dienų technologijoms.

BDAR įsigaliojimas įmonėms reiškė didesnes išlaidas ar sunkumus vystant savo produktus. Kita vertus, BDAR sukūrė įmonėms galimybes tobulėti kibernetinio saugumo ir kitose srityse.

3. Duomenų anonimizacija

Vienas iš būdų spręsti su BDAR susijusius duomenų saugojimo iššūkius yra duomenų anonimizacija. Anonimizacijos proceso metu duomenų rinkinio atributai dažniausiai skirstomi į atskiras keturias grupes [GMV+18]:

- **Tikslūs identifikatoriai**. Tai tokie duomenys, kuriuos galima susieti su vienu asmeniu, pavyzdžiui, el. pašto adresas ar asmens kodas.
- **Beveik tikslūs identifikatoriai**. Tai tokie duomenys, kurie tiesiogiai nėra susiję su asmeniu, bet, sujungus kelis požymius, iš jų galima atpažinti asmenį, pavyzdžiui, gimimo data, pašto kodas, specialybė.
- **Neskelbtina informacija**. Tai tokie duomenys, kurių asmenys nenori atskleisti ar būti susieti su jais. Pavyzdžiui, ligos, finansinė padėtis.
- **Nejautri informacija**. Tai duomenys, kurie nepatenka į nei vieną iš anksčiau išvardintų kategorijų.

Visgi, šis skaidymo būdas ne visada yra efektyvus, nes, pavyzdžiui, reta liga gali atskleisti tikslų žmogų. Be to, pačiose kategorijose yra didelių skirtumų: gyvenamoji vieta gali reikšti milijonus žmonių didelio miesto atveju arba tik mažą žmonių grupę kaimo atveju. Anonimizacija yra procesas, kurio metu slepiamas ryšys tarp duomenų ir kam jie priklauso [DS16]. Šis procesas susideda iš dviejų pagrindinių etapų:

- **De-identifikacijos**. Šio etapo metu slopinami tikslūs identifikatoriai.
- **Beveik tikslių identifikatorių maskavimas**. Kadangi beveik tikslūs identifikatoriai turi analitinę vertę, jie yra maskuojami, o ne slopinami.

Anonimizacijos metodai dažnai naudojami siekiant išsaugoti privatumą, daugiausia dėmesio skiriant asmens duomenų pavertimui į anoniminius duomenis, siekiant sumažinti neteisėto atskleidimo riziką [MBR+19]. Tačiau, padidėjus anonimiškumui, duomenų tikslumas (ir jų nauda) sumažėja. Taigi, organizacija turi pasirinkti tinkamus anonimizacijos metodus. Murthy ir kt. išskiria 5 duomenų anonimizacijos būdus. Išskiriant juos naudojama pavyzdinė lentelė, kuri turi ID, vardo, adreso, pašto kodo ir elektros suvartojimo (kW) stulpelius. Darbe aprašomi šie metodai:

1. **Apibendrinimas**. Tai procesas, kurio metu įrašas pakeičiamas ne tokia tikslia reikšme. Ši technika pritaikoma įrašo vienam iš stulpelių (angl. cell level) ir prideda jam netvarkos. Visgi, apibendrinimas negali būti taikomas visiems atributams. Šis metodas dažniausiai taikomas beveik tiksliems identifikatoriams, pavyzdžiui, gimimo datą keičiant į amžių, o vėliau amžių keičiant į amžiaus rėžį. Naudojantis lentelės duomenimis, tokie atributai kaip ID ar vardas negali būti apibendrinami. Adresą galima apibendrinti panaikinant namo numerį ir

paliekant tik gatvę, o vietoj tikslaus elektros suvartojimo galima jį apibendrinti į rėžius, pavyzdžiui, 434 pakeičiant į 400–450.

2. **Slopinimas**. Šio proceso metu visa duomenų įrašo dalis yra panaikinama, pakeičiant vertę į tokią, kuri neturi reikšmės, pavyzdžiui „***“. Ši technika paslepia informaciją ištrindama ją, o pradiniuose duomenyse esantys įrašai visiškai pašalinami iš galutinės išvesties. Pritaikius šią techniką duomenys gali tapti nebenaudojami ir turėti neigiamą įtaką tyrimams. Šis metodas dažniausiai taikomas tiksliems identifikatoriams. Straipsnio autoriai teigia, kad šią techniką galima naudoti norint paslėpti ID ar vardą (duomenys, kurie nėra svarbūs tyrimui), bet negalima ja naudojantis anonimizuoti adreso ar pašto kodo.
3. **Iškraipymas**. Šiuo proceso metu duomenys yra pakeičiami kitais, bet vėliau juos galima atstatyti į pradinius.
4. **Sukeitimas**. Šio proceso metu įrašai yra atsitiktinai pertvarkomi. Šis būdas gali būti naudojamas norint surinkti to paties atributo duomenis. Visgi, šio metodo negalima naudoti visiems duomenims, nes tyrimo rezultatai gali būti netikslūs. Be to, išlieka tikimybė, kad po atsitiktinio sukeitimo naujieji duomenys atitiks pradinius. Darbo autoriai pateikia pavyzdį, kuriame atsitiktine tvarka yra sukeičiami vardai pradiniuose duomenyse.
5. **Maskavimas**. Tai technika, kurios metu pasirinkto atributo simboliai keičiami į kitus. Bet koks skaitmuo bus pakeistas į 1, bet kokia mažoji raidė į z, didžioji raidė į Z. Pirmasis įrašo simbolis, 0 ir specialieji simboliai nebus maskuojami. Šis metodas naudoja daug resursų tikrinant ir keičiant simbolius, bet duomenys tampa nebenaudojami. Vietoj maskavimo galima naudoti slopinimą, pakeičiant kai kuriuos simbolius, taip pasiekiant tuos pačius rezultatus kaip maskavimas, bet su didesniu efektyvumu.

Taip pat išskiriamos ir šios anonimizacijos technikos [NKU18]:

1. **Visos srities apibendrinimas**. Ši technika pasižymi dideliu duomenų iškraipymu. Naudojant ją duomenis yra apibendrinami iki tam tikro hierarchijos lygio. Pavyzdžiui, visų sričių mokytojai yra žymimi mokytojais, o kiti mokyklos darbuotojai – ne mokytojais.
2. **Pomedžio apibendrinimas**. Šiuo atveju visos viršūnės (angl. node) tame pačiame lygyje yra apibendrinamos iki tokio pat lygio arba nėra anonimizuojami iš viso.
3. **Langelio (angl. cell) apibendrinimas**. Ši technika skiriasi nuo aukščiau išvardintų tuo, kad ji taikoma tik vienam įrašui.
4. **Daugialypis apibendrinimas**. Ši apibendrinimo technika skirta apibendrinti kelių atributų kombinacijas.
5. ***Bucketization***. Ši technika skirsto duomenis į atskiras grupes ir tuomet atskiria beveik tikslius identifikatorius nuo jautrios informacijos.

Kadangi turint beveik tikslių identifikatorių rinkinį [GKK+07] galima nustatyti tikslų žmogų, buvo pasiūlytas *k*–anonimiškumo modelis: kiekvienam įrašui lentelėje turi būti bent $k - 1$ įrašas su tokiu pačiu beveik tikslių identifikatorių rinkiniu. Įrašai su identiškais beveik tiksliais identifikatoriais sudaro ekvivalentišką klasę. *k*–anonimiškumas dažniausiai pasiekiamas naudojant apibendrinimą arba slopinimą, o tai veda prie informacijos praradimo. *k* – anonimiškumas užtikrina, kad tikimybė atpažinti tikslų asmenį duomenų rinkinyje yra ne didesnė nei $1/k$ [NKU18].

Norint pašalinti *k*–anonimiškumo trūkumus buvo pristatyta *l* – įvairovės sąvoka [GKK+07], nes *k*–anonimiškumas gali atskleisti daug jautrios informacijos, kai yra daug identiškos jautrios informacijos ekvivalentumo klasėje (pavyzdžiui, visi žmonės serga ta pačia liga). *l* – įvairovė neleidžia atlikti vienodumo ir bendrų žinių išpuolių užtikrinant, kad bent jau *l* jautrios reikšmės yra teisingai išreikštos kiekvienoje ekvivalentumo klasėje (pavyzdžiui, tikimybė susieti duomenų aibę (angl. tuple) su jautria informacija yra apribota iki $1/l$. Bet koks *k* – anonimiškumo algoritmas gali būti pritaikytas pasiekti *l* – įvairovę pakeičiant ekvivalentiškumo klasės validavimo sąlygą.

Dar vienas anonimiškumo modelis yra *t* – artumo [NKU18]. Jį įgyvendinus bet kurioje jautrių atributų grupėje pasiskirstymas, identifikuojant atributą, turi būti artimas atributo pasiskirstymui visoje lentelėje.

Nayahi ir kt. [NK17] išskiria anonimizuotų duomenų pažeidžiamumą įvairioms atakoms:

1. **Homogeniškumo ataka**. Kai visi įrašai anonimizuotame rinkinyje turi tokias pačias jautrių atributų reikšmes, *k* – anonimizacija praranda prasmę. *l* – įvairovės modelis padeda išvengti šios atakos.
2. **Panašumo ataka**. Net jeigu jautrių atributų reikšmės nėra tokios pačios, jos gali būti panašios. Nepaisant to, kad anonimizuotam rinkiniui yra pritaikytas *l* – įvairovės modelis, jautrių atributų reikšmės gali būti atskleistos.
3. **Bendrųjų žinių ataka**. Užpuolikas, turintis tam tikrą pagrindinę informaciją apie jautrius požymius arba apie bendras duomenų charakteristikas, naudoja ją, kad sumažintų galimų neskelbtinų reikšmių neekvivalenčių klasės grupių skaičių.
4. **Tikimybės išvados ataka**. Šios atakos grėsmė atsiranda, kai viena jautraus atributo reikšmė pasikartoja dažniau negu kitos ekvivalentiškumo klasėje.

Karle ir Vora [KV17] atliko *Datafly* ir *Mondrian* algoritmų palyginimą. Darbo autoriai teigia, kad *Datafly* yra gobšus (angl. greedy) euristinis vienmatis algoritmas, atliekantis visų duomenų apibendrinimą. Algoritmas skaičiuoja beveik tikslių identifikatorių dažnį aibėje ir, jei *k* – anonimiškumas dar nėra pasiektas, algoritmas apibendrina labiausiai išsiskiriančias reikšmes. *Datafly* atlieka visų duomenų apibendrinimą, kol kiekviena beveik tikslių identifikatorių kombinacija pasikartoja bent jau *k* kartų. Šis algoritmas geriausiai tinka sintetinių duomenų anonimizacijai.

Mondrian algoritmas yra gobšus daugiamatis algoritmas, kuris padalija duomenis rekursyviai į kelias skirtingas grupes, kiekvienoje iš jų esant bent k įrašų. Algoritmas pradeda anonimizaciją nuo labiausiai apibendrintos atributo reikšmės beveik tikslių identifikatorių rinkinyje. Teigiama, kad šis algoritmas labiausiai tinka realiam duomenų rinkiniui.

Eyupoglu ir kt. [EAZ+18] išskiria 8 žingsnius algoritmui, kuriame išsaugomi duomenų nauda ir privatumas:

1. Nurodomas pradinis duomenų rinkinys, beveik tikslūs atributai ir neskelbtinos informacijos atributai.
2. Randamos unikalios kiekvieno beveik tikslaus atributo reikšmės.
3. Kiekvienam beveik tiksliam identifikatoriui apskaičiuojamas įrašų kiekis, kuriame yra unikalių atributo reikšmių.
4. Unikalios reikšmės yra surūšiuojamos didėjančia tvarka pagal dažnį.
5. Pradiniame duomenų rinkinyje randamos unikalių atributų vietos tolimesniems atsitiktinumo ir pakeitimo procesams.
6. Apskaičiuojamas kritinis unikalių atributų reikšmių kiekis kiekvienam beveik tiksliam identifikatoriui.
7. Apskaičiuojamos naujos reikšmės visiems pasirinktiems kritiniams unikaliems atributams.
8. Atnaujinamas pradinis duomenų rinkinys.

Sukurtas algoritmas buvo testuojamas naudojant 1994 metų JAV surašymo duomenis. Beveik tiksliais identifikatoriais pasirinkti amžius, rasė ir lytis, o neskelbtina informacija – atlyginimas. Tyrimo rezultatai rodo, kad siūlomas algoritmas tik šiek tiek iškreipia pradinius duomenis ir yra efektyvesnis už kitus, lyginime naudotus algoritmus. Darbo autoriai teigia, kad algoritmas užtikrina asmeninių duomenų saugumą prieš paviešinant duomenis.

Duomenys gali būti saugomi ir grafuose. Išskiriami 6 grafo duomenų (angl. graph data) anonimizacijos būdai [JLM+15]:

1. **ID pašalinimas**. Šis metodas yra labai populiarus, nepaisant to, kad jis yra itin pažeidžiamas deanonimizacijos atakų. Dažnai naudojamas dėl savo paprastumo, lengvo pritaikymo.
2. **Briaunų redagavimu pagrįsta anonimizacija**. Norint išsaugoti grafo duomenų privatumą buvo pristatytos dvi briaunų redagavimo schemos: pridėti/atimti (angl. Add/Del) ir sukeitimo (angl. Switch). Pirmuoju atveju, k atsitiktinių briaunų bus pridėta ir k atsitiktinai pasirinktų briaunų bus ištrinta. Sukeitimo schemoje yra atliekami k atsitiktinių briaunų sukeitimai.
3. k–**anonimiškumas**. Šis modelis dažnai naudojamas anonimizuojant duomenis, todėl buvo bandoma pritaikyti jį ir grafams.

4. **Agregacija/Klasėmis/Grupėmis pagrįsta anonimizacija**. Norint apsaugoti duomenis, jie yra sugrupuojami. Vienas iš minimų algoritmų suskirsto vartotojus ir tuomet aprašo grafą suskirstymo (angl. partition) lygmenyje.
5. **Diferencialinis privatumas**. Darbo autoriai teigia, kad tai labai populiarus metodas. Iš pradžių jis buvo pristatytas statistinėms duomenų bazėms, bet pastaruoju metu bandoma šį metodą pritaikyti ir grafams. Jis stengiasi užtikrinti grafo briaunų privatumą.
6. **Atsitiktiniu keliu pagrįsta anonimizacija**. Grafo briaunos yra pakeičiamos atsitiktinio kelio seka.

Darbo autoriai pristatė grafo duomenų anonimizacijos algoritmą *SecGraph*. Jis susideda iš 3 modulių:

1. **Anonimizacijos modulis**. Šio modulio tikslas yra anonimizuoti pradinius grafo duomenis ir sugeneruoti anonimizuotus. Šiame modelyje naudojamos 11 grafo duomenų anonimizacijos schemų, kurios pritaiko aukščiau išvardytus būdus.
2. **Naudingumo modulis**. Šiame modulyje yra įvertinamas pradinių ir anonimizuotų duomenų naudingumas. Šiame modulyje tikrinama ar anonimizuoti duomenys gali būti paviešinti.
3. **Deanoninimizacijos modulis**. Šiame modulyje yra tikrinamas duomenų saugumas. Jame naudojami 15 deanonimizacijos algoritmų, kuriais patikrinama, ar anonimizuoti duomenys yra atsparūs šiuolaikiniams deanonimizacijos algoritmams.

Tyrimo rezultatai rodo, kad anonimizuoti grafo duomenys išlieka pažeidžiami deanonimizacijos atakoms. Pažeidžiamumo lygis priklauso nuo to, kiek naudingumo išlieka anonimizuotose duomenyse.

Sei ir kt. [SOT+17] pristatė anonimizacijos algoritmą, skirtą jautrių beveik tikslių identifikatorių anonimizacijai. Algoritmas susideda iš dviejų dalių: randomizacijos, atliekamos duomenų valdytojo, ir atstatymo, atliekamo duomenų analitiko. Pagal duomenų valdytojo pasirinktus parametrus algoritmas sukuria anonimizuotą įrašą „agreguotoje išraiškoje“ ir įterpia jį į anonimizuotą duomenų bazę. Duomenų analitikas pirmiausia nustato, kurie beveik tikslūs identifikatoriai turėtų būti analizuojami. Tada įvertinamas įrašų skaičius kiekvienam pasirinktų beveik tikslių identifikatorių verčių deriniui. Darbo autoriai pateikia pavyzdį: duomenų bazė turi vieną įrašą, kuriame yra žmogaus vardas (Alisa), amžius (41), adresas (10105) bei liga (ŽIV). Vartotojui pasirinkus, kad tris kintamieji yra lygūs 2, bus sugeneruoti 6 įrašai, o vienas iš jų atitiks tikrąjį įrašą apie Alisą. Tokiu atveju, net jeigu duomenų analitikas žinotų 2 iš 3 beveik tikslių identifikatorių, jis negalėtų nustatyti trečiojo beveik tikslaus identifikatoriaus reikšmės su tikimybe didesne nei 1/2. Sukurtas algoritmas buvo testuojamas naudojant JAV surašymo duomenis. Tyrimo rezultatai rodo, kad siūlomas metodas gali anonimizuoti ir atstatyti duomenų bazes išlaikant aukštą duomenų kokybę.

Liu ir kt. [LPC+18] teigia, kad klasikiniai anonimizacijos algoritmai duomenų apdorojimo metu panaikina nepilnus įrašus medicininių duomenų rinkiniuose. Norėdami pasiūlyti šios problemos sprendimą, jie pristatė anonimizacijos algoritmą, kuris skirtas nepilnam medicininių duomenų rinkiniui. Šis algoritmas atitinka l – įvairovės modelį, o jo veikimas yra pagrįstas sankaupomis (angl. clustering). Norint išsaugoti įrašus, kuriuose trūksta duomenų, algoritmas skirsto duomenis ir naudoja informacijos apibendrinimo generuotą informacijos entropiją, kad apskaičiuotų atstumą sankaupos etape. Tuomet apibendrinamos sankaupos būdu gautos duomenų grupės. Algoritmas buvo testuojamas naudojant suaugusiųjų medicininių duomenų rinkinį su 32561 įrašais ir 15 skirtingų atributų. Tyrimo rezultatai rodo, kad jis gali geriau apsaugoti jautrius pacientų duomenis, sumažinti informacijos praradimą anonimizuojant trūkstamus duomenis ir pagerinti duomenų rinkinio prieinamumą.

Ciglic ir kt. [CEK16] tyrė duomenų rinkinių, kurie turi nulines reikšmes (angl. null values), anonimizaciją. Darbo autoriai teigia, kad nulinės reikšmės dažnai pasitaiko duomenų rinkiniuose. Tačiau dabartiniai anonimizacijos algoritmai reikalauja, kad anonimizuojamas duomenų rinkinys neturėtų įrašų su nulinėmis reikšmėmis. Tyrėjai išskiria 3 būdus, kaip galima lyginti nulines reikšmes:

- Įprastas palyginimas (angl. basic match): nulinė reikšmė nėra lygi nei kitai nulinei reikšmei, nei kitoms reikšmėms.
- Išplėstas palyginimas (angl. extendend match): nulinė reikšmė lygi kitai nulinei reikšmei.
- Galimas palyginimas (angl. maybe match): nulinė reikšmė lygi bet kuriai kitai reikšmei.

Tyrimo metu buvo buvo naudojama patobulinta apibendrinimo anonimizacijos technika, kuri įtraukia ir nulines reikšmes. Ji buvo įgyvendinta *Java* programavimo kalba. Nulinės reikšmės buvo apdorojamos dviem būdais: panaikinant visus įrašus su nulinėmis reikšmėmis prieš anonimizacijos procesą (įprastas palyginimas) ir vertinant nulines reikšmes kaip visas kitas (išplėstas palyginimas). Algoritmas susideda iš 2 dalių: lentelės paieškos ir privatumo testo. Ieškodamas optimalaus sprendimo algoritmas naudoja *best-first* paieškos algoritmą. Sukurtas algoritmas buvo testuojamas naudojant 1994 metų JAV surašymo duomenis su 45222 įrašais. Tyrimo rezultatai rodo, kad įtraukiant nulines reikšmes į anonimizacijos algoritmą yra prarandama mažiau informacijos.

Domingo ir Soria [DS16] tyrė anonimizacijos galimybes didžiuosiuose duomenyse. Tyrimo metu daugiausia dėmesio buvo telkiama k – anonimiškumui užtikrinti. Darbo autoriai išskiria du pagrindinius iššūkius norint naudoti k – anonimiškumą didžiuosiuose duomenyse:

- Neskelbtina informacija gali būti laikoma beveik tiksliais identifikatoriais, o tai gali privesti prie didelio informacijos kiekio praradimo norint atitikti k – anonimiškumą.
- Didžiuosiuose duomenyse gali būti neribotas kiekis duomenų valdytojų, kurie gali paviešinti

nepriklausomus duomenų rinkinius apie sutampančias tiriamųjų grupes.

Darbo autoriai atsižvelgdami į aukščiau išvardytus iššūkius, siūlo keisti neskelbtinos informacijos apibendrinimą į alternatyvią agregacijos techniką, kuri yra mažiau jautri komponentų pokyčiams k – anonimiškumo klasėje. Siūloma naudoti arba jautrios informacijos atributų vidurkį, arba jų medianą. Darbo autoriai teigia, kad su šiais pakeitimas, k – anonimiškumo pritaikymas išsaugo savo naudingumą didžiuosiuose duomenyse.

4. Tyrimas

Siekiant išnagrinėti galimybę pritaikyti duomenų anonimizaciją įgalinant BDAR duomenų modeliuose buvo atliktas tyrimas. Jo metu:

- Pasirinktas duomenų modelis užpildytas generuotais duomenimis.
- Įgyvendintas ir pasirinktam duomenų modeliui pritaikytas duomenų anonimizacijos rinkinys.
- Ištirtas anonimizacijos poveikis pasirinkto modelio duomenims.

Tyrimas buvo atliekamas naudojant *PostgreSQL* reliacinę duomenų bazių valdymo sistemą.

4.1. Tyrimo eiga

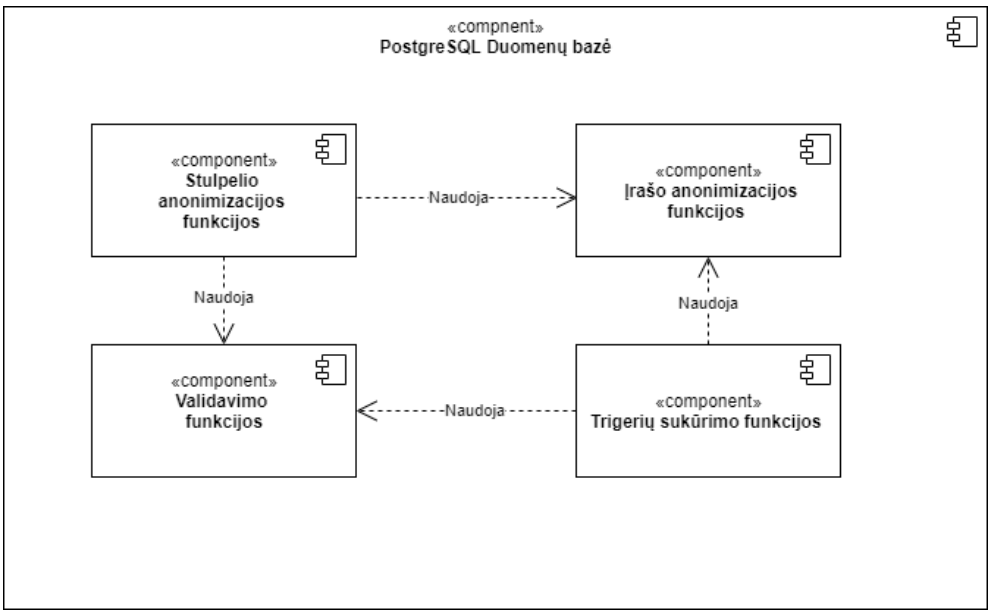
4.1.1. Projektavimas

Tyrimo pirmajame etape išskirti šie duomenų anonimizacijos metodai, kurie bus įgyvendinami prototipe:

1. **Apibendrinimas**. Bus įgyvendinti skirtingiems duomenų tipams tinkantys variantai: tekstinėms (panaikinant visus skaičius iš reikšmės, panaikinant paskutinius *n* reikšmės simbolių), skaitinėms (pakeičiant reikšmę į rėžį, pridedant netvarkos), datoms (datą pakeičiant į tų metų sausio 1, pakeičiant datą į amžių, pakeičiant datą į amžiaus rėžį).
2. **Sukeitimas**. Šio metodo atveju pradiniai duomenys bus sukeičiami atsitiktine tvarka. Nurodytame stulpelyje atliekamas atsitiktinis duomenų sukeitimas. Pavyzdžiui, vieno žmogaus el. pašto adresas būtų priskirtas kitam ir t.t.
3. **Slopinimas**. Šio metodo atveju pradinė duomenų vertė bus pakeičiama į neturinčią jokios reikšmės. Toks metodas tiktų anonimizuojant vardą, pradinę reikšmę pakeičiant į „***“.
4. **Iškraipymas**. Bus įgyvendinti 2 skirtingiems duomenų tipams tinkantys variantai: sugeneruojant atsitiktinius skaitinius ir tekstinius duomenis

4.1.2. Programavimas

Tyrimo antrajame etape naudojant *PG/pgSQL* programavimo kalbą buvo sukurtas duomenų anonimizacijos prototipas. Jo metu sukurtas funkcijas galima išskirti į 4 grupes: vieno įrašo lygyje veikiančios anonimizacijos funkcijas, pagalbines (skirtos validavimui), stulpelių anonimizacijos ir funkcijas skirtos trigerių sukūrimui. 1 pav. pavaizduota sukurto prototipo funkcijų sąsajos. Stulpelio anonimizacijos ir trigerių sukūrimo komponentai naudoja validavimo ir įrašo anonimizacijos komponentus.



1 pav. Prototipo komponentų diagrama

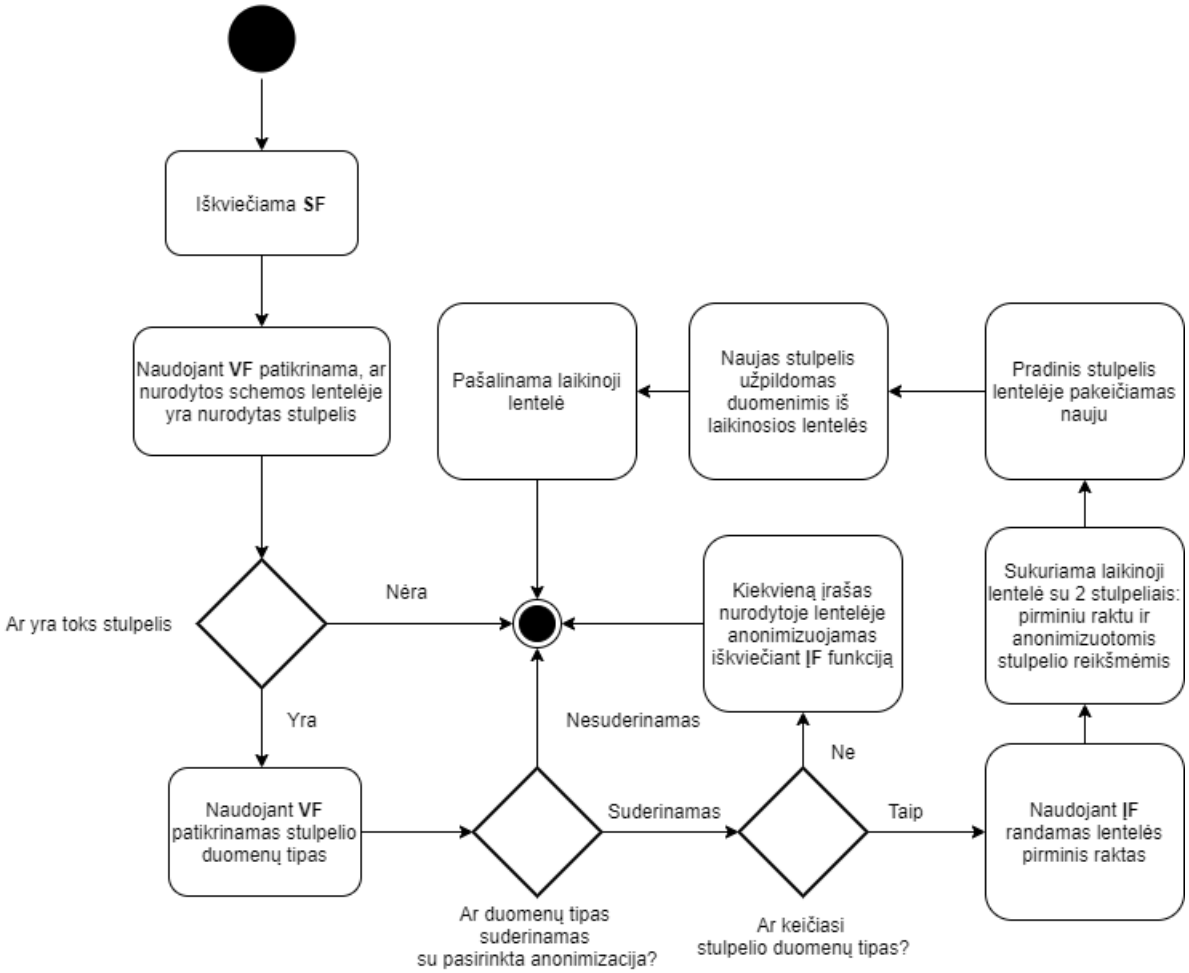
PostgreSQL reliacinėje duomenų bazių valdymo sistemoje *information_schema* susideda iš virtualių lentelių rinkinio, kuriame saugoma informacija apie duomenų bazės objektus [Gro]. Naudojantis šiomis virtualiomis lentelėmis galima, pavyzdžiui, rasti informaciją apie lenteles, stulpelius ar ribojimus (angl. constraints). Tai suteikia galimybę atlikti tokius veiksmus, kaip stulpelio duomenų tipo patikrinimą ar lentelės pirminio rakto radimą. Validavimo funkcijos naudoja šį virtualių lentelių rinkinį. Validavimo funkcijos (žemiau esančiose diagramose trumpinamos kaip **VF**) skirtos šiems veiksmams:

1. Naudojant stulpelio, lentelės ir schemos pavadinimus patikrinti, ar egzistuoja toks stulpelis nurodytoje lentelėje, kuri yra nurodytoje schemoje, naudojant *information_schema.columns* lentelę, *table_name*, *schema_name* ir *column_name* stulpelius.
2. Patikrinti pateikto stulpelio duomenų tipą. Tam naudojama *information_schema.columns* lentelę, *data_type* stulpelis.
3. Rasti lentelės pirminį raktą. Tam naudojamos *information_schema.table_constraints* ir *information_schema.key_columns_usage* lentelės ir *information_schema.key_column_usage* lentelės *column_name* stulpelis.

Vieno įrašo anonimizacijos funkcijos (žemiau esančiose diagramose trumpinamos kaip **ĮF**) skirtos anonimizuoti jai perduotą vieną reikšmę. Sukurtos funkcijos skirtos panaikinti visus skaitmenis iš tekstinės reikšmės, pakeisti paskutinius *n* simbolių žvaigždute, pridėti netvarkos pagal nurodytą lygį pateiktam skaičiui, pateiktą skaičių pakeisti į rėžį, apskaičiuoti amžių pagal pateiktą datą, pateiktos datos mėnesį ir dieną pakeisti sausio 1.

Sukurtos stulpelių anonimizacijos (žemiau esančiose diagramose trumpinamos kaip **SF**)

funkcijos aprašomos 4.1.1 poskyryje. Jų veikimas bendruoju atveju pavaizduotas 2 pav. veiklos diagrama. Kiekviena sukurta funkcija priima 3 parametrus: schemos, lentelės ir stulpelio pavadinimus. Dalis funkcijų priima papildomus parametrus, pavyzdžiui, anonimizuojant skaitinę reikšmę pasirenkama, iš kokio skaičiaus dauginama atsitiktinė reikšmė arba netvarkos lygis. Pirmuoju žingsniu iškvietus validavimo funkciją patikrinama, ar yra toks stulpelis nurodytoje lentelėje, kuri yra nurodytoje schemoje. Tuomet, naudojant kitą validavimo funkciją patikrinama, ar duomenų tipas yra suderinamas pasirinkta anonimizacija. Jeigu duomenų tipai suderinami ir nesikeičia stulpelio duomenų tipas, tuomet naudojant vieno įrašo anonimizacijos funkciją anonimizuojamas kiekvienas įrašas nurodytoje lentelėje. Jeigu reikia keisti stulpelio duomenų tipą, tuomet naudojant validavimo funkciją randamas lentelės pirminis raktas ir naujai sukurta lentelė užpildoma 2 stulpeliais: pirminiu raktu ir anonimizuotomis stulpelio reikšmėmis. Vietoje seno stulpelio sukurtas naujas yra užpildomas duomenimis iš laikinosios lentelės. Paskutinis žingsnis yra laikinosios lentelės pašalinimas (angl. drop table).

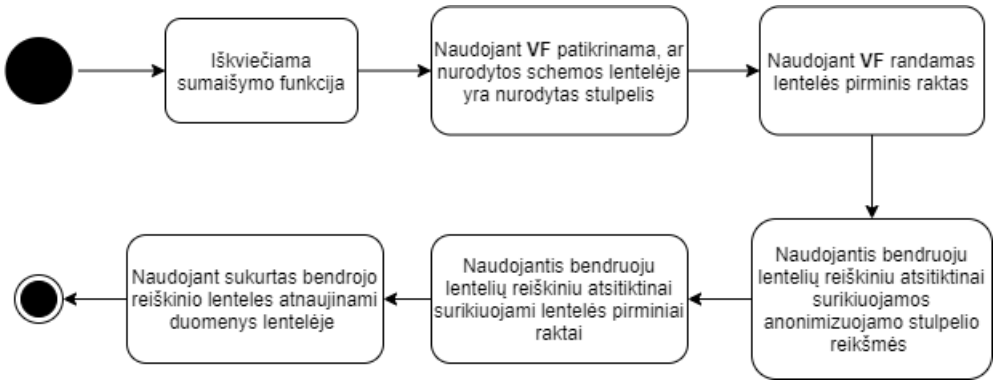


2 pav. Stulpelio anonimizacijos funkcijų veikimas

Sukeitimo funkcijos veikimas pateikiamas 3 pav.. Ši funkcija priima 3 parametrus: schemos,

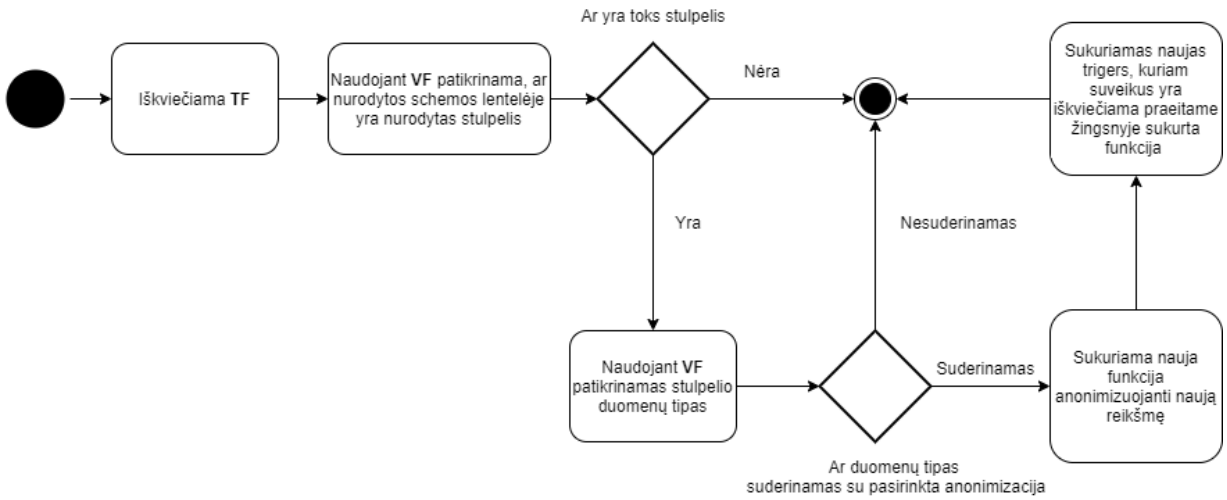
lentelės ir stulpelio pavadinimus. Iš pradžių atliekama validacija, ar egzistuoja stulpelis ir randamas pirminis lentelės raktas. Tuomet naudojantis dviem bendraisiais lentelių reiškiniais (angl. Common Table Expression) duomenys yra atsitiktinai sukeičiami. Pirmojoje lentelėje atsitiktine tvarka surūšiuojami anonimizuojamo stulpelio įrašai ir jiems priskiriamas eilės numeris. Antrojoje lentelėje atsitiktine tvarka surūšiuojami pirminiai raktai ir jiems priskiriamas eilės numeris. Tuomet atnaujinami anonimizuojamos lentelės duomenys: kiekvieno įrašo stulpeliui priskiriama nauja reikšmė. Žemiau pavaizduotas sukeitimo algoritmas, jame – *CTE1* ir *CTE2* bendrieji lentelių reiškiniai, *column* anonimizuojamos lentelės *Table* stulpelis, *primary_key*– pirminis lentelės raktas.

```
WITH CTE1 AS
(SELECT row_number() OVER (ORDER BY random()) n,
column AS i1 FROM Table),
CTE2 AS
(SELECT row_number() OVER (ORDER BY random()) n,
primary_key AS i2 FROM Table)
UPDATE Table SET column_name = CTE1.i1
FROM CTE1 JOIN CTE2
on CTE1.n = CTE2.n WHERE primary_key = CTE2.i2;
```



3 pav. Stulpelio anonimizacijos funkcijos veikimas

Sukurtos trigerių funkcijos (žemiau esančiose diagramose trumpinamos kaip **TF**) leidžia užtikrinti, kad naujai įterpiami arba atnaujinami duomenys būtų taip pat anonimizuoti. Jų veikimas pavaizduotas 4 pav.. Šios funkcijos priima šiuos parametrus: stulpelio, lentelės, schemos, norimų sukurti trigerių ir funkcijos pavadinimus bei trigerio tipą. Kaip ir stulpelio anonimizacijos atveju, visų pirma, atliekamas validavimas. Tuomet sukuriama nauja funkcija, kurį modifikuoja naują reikšmę, kad ji būtų anonimizuota. Tuomet sukuriamas naujas trigeris, kuris suveikia įterpiant arba atnaujinant duomenis.



4 pav. Trigerio ir funkcijos sukūrimas

4.1.3. Duomenų generavimas

Sintetinių duomenų rinkinių generavimas yra įprasta praktika daugelyje tyrimų sričių [ALM11]. Tokie duomenys dažnai generuojami tenkinant specifinius poreikius ar tam tikras sąlygas, kurias tikruose duomenyse gali būti sunku rasti. Duomenų pobūdis skiriasi priklausomai nuo taikymo srities ir apima tekstinius, grafinius ir kitus duomenis. Įprastas tokių sintetinių duomenų rinkinių kūrimo procesas yra mažų scenarijų (angl. script) ar programų įgyvendinimas, apsiribojant nedidelėmis problemomis ar konkrečia programa. Prieiga prie tikrų mikroduomenų yra labai apribota, o viešai prieinama paprastai yra anonimizuota arba apibendrinta. Todėl Ayala-Rivera ir kt. [APM+16] sukūrė tikroviškų sintetinių mikroduomenų generavimo sistemą (COCOA), leidžiančią apibrėžti daugelio atributų ryšius, siekiant išsaugoti duomenų funkcines priklausomybes. Atsižvelgiant į norimų duomenų savybes, COCOA gali efektyviai sukurti daugialypius duomenų rinkinius. Tyrimas parodė, kad svarbu naudoti išsamų įvairių bandymo duomenų rinkinių rinkinį ir kad COCOA yra efektyvus skaičiavimo resursų atžvilgiu, todėl jį galima naudoti ir realiems projektams. Taip pat, tyrimas parodė, kad COCOA yra naudinga testuojant anonimizacijos metodus didinant testavimo scenarijų kiekį ir įvairovę.

Šio tyrimo metu duomenų generavimui naudota *.NET* biblioteka *Bogus* [Cha]. Duomenų modelis remiasi *Pagilla* duomenų baze [Web]. Tai DVD nuomos duomenų modelis, kurį sudaro 15 lentelių. Jame atlikti du pakeitimai: *Staff* lentelėje pridėti algos ir gimimo metų stulpeliai. Generuojant duomenis naudotos 4 lentelės: *Customer*, *Staff*, *Address* ir *City*. Tyrimui buvo sugeneruoti 100 tūkst. *Customer*, 1000 *Staff*, 90 tūkst. *Address*, 1000 *City* lentelių įrašai. Duomenys įterpti į PostgreSQL duomenų bazę naudojant *Entity Framework Core* biblioteką. Taikytas *Database First* metodas, kai *C#* klasės sugeneruojamos pagal duomenų bazės lentelės ir jų ryšius. Generuojant

duomenis buvo aprašomos taisyklės kiekvienai klasės savybei (angl. property). Pavyzdžiui, vartotojo vardas arba el. pašto adresas generuojami naudojant vardą ir pavardę (pavyzdžiui, Aric Murray sugeneruotas el. pašto adresas Aric45@gmail.com, o Veda Nader – Veda_Nader29@gmail.com). Sukurtoje programoje duomenys gali būti terpiami į vieną iš 4 lentelių nepriklausomai viena nuo kitos, pasirenkant, į kurias lenteles duomenis bus terpiami.

4.1.4. Testavimas

Pasirinktame modelyje asmeniniais duomenimis galima laikyti šiuos: *Customer* lentelėje esančius vardą, pavardę, el. pašto adresą, *Address* lentelėje esančius pašto kodą, adresą, telefono numerį, *Staff* lentelėje esančius vardą, pavardę, gimimo datą, algą, el. pašto adresą, naudotojo vardą, slaptažodį. Testuojant prototipą buvo naudojami paminėtų trijų lentelių duomenys. Testavimas vyko keturiais etapais: validavimo funkcijų testavimas, vieno įrašo lygyje veikiančių funkcijų testavimas, stulpelio anonimizacijos funkcijų testavimas, funkcijų, skirtų trigerių sukūrimui, testavimas.

Validavimo funkcijos testuotos tikrinant jų grąžinamas reikšmes. Pirminio rakto paieškos funkcija tikrinta ją iškviečiant su lentelės, kurios pirminis raktas yra žinomas, pavadinimu, neegzistuojančios lentelės pavadinimu, su tuščia eilute ir lentelės, kurios pirminis raktas yra sudėtinis, pavadinimu. Stulpelio duomenų tipo paieškos funkcija testuota iškvietus ją su žinomo duomenų tipo stulpelio pavadinimu, tuščia eilute, neegzistuojančio stulpelio pavadinimu. Funkcija, kuri patikrina, ar egzistuoja stulpelis nurodytoje lentelėje ir schemoje, testuota su neegzistuojančios lentelės ir/ar stulpelio pavadinimais ir su egzistuojančiais parametrais.

Vieno įrašo lygyje veikiančios funkcijos testuotos tikrinant, ar jos veikia korektiškai iškviečiant jas su įvairiais parametrais. Tikrinta, ar funkcija, skirta visų skaitmenų iš teksto panaikinimui, veikia, kai tekstas turi ir neturi savyje skaitmenų, kai susideda tik iš skaitmenų, kai yra tuščias. Tikrintas funkcijos, pridedančios netvarkos skaičiams, veikimas, jai perdavus įvairias reikšmes, testavimo metu funkcijoje pridėtas patikrinimas, ar netvarkos lygio parametras yra tarp 1 – 100. Tikrinta, ar teisingai konvertuojama data į amžių, data į pirmą datos metų dieną. Tikrinta, ar teisingai pakeičiami paskutiniai *n* teksto simboliu žvaigždutėmis, su parametrais *n=0*, *n > 0 ir n < simbolių skaičių*, *n > simbolių skaičių*. Tikrinta, ar teisingai konvertuojamas skaičius į rėžį.

Vieno įrašo lygyje veikiančios funkcijos buvo testuotos ir kuriant materializuotas virtualiąsias lenteles. Pirmojoje sukurtoje materializuotoje virtualioje lentelėje atvaizduoti klientų ID, vardai, pavardės (*Customer* lentelė), adresai ir pašto kodai (*Address* lentelė). Vardo ir pavardės stulpeliams pritaikytas slopinimas, adresui ir pašto kodui apibendrinimas. Šiame žingsnyje rasta klaida:

nebuvo atsižvelgta, kad tekstinė reikšmė gali turėti ’ simbolį. Klaida pataisyta reikšmėje pakeičiant kiekvieną ’ į ”. Pradiniai duomenys, pavyzdžiui, *{1, Maria, Smith, 1913 Hanoi Way, 35200}*, yra pakeičiami į *{1, ***, ***, **** Hanoi Way, 352**}*. Vietoje tikslaus gatvės numerio ir gatvės pavadinimo anonimizuotose duomenyse lieka tik gatvės pavadinimas bei pašto kodo dalis.

Antrojoje materializuotoje virtualioje lentelėje buvo atvaizduoti darbuotojų ID, vartotojų vardai, algos, gimimo datos (*Staff* lentelė) ir jų telefono numeriai (*Address* lentelė). Vartotojo vardui pritaikytas iškraipymas, algai apibendrinimas iki rėžių, gimimo datai apibendrinimas iki amžiaus, telefono numeriui apibendrinimas. Pradiniai duomenys, pavyzdžiui, *{3, Eduardo_Watsica, 1184, 1994-03-25, 4963224925}*, yra pakeičiami į *{3, c90b71e413e56e452cc1a4ac5b41315f, [1000,1500), 26, 496322****}*. Vietoje tikslios algos, gimimo datos ir telefono numerio lieka apibendrinti duomenys: algos rėžis, amžius bei dalis telefono numerio.

Testuojant sukeitimo funkciją buvo sukurta laikinoji lentelė, nes virtualiosios lentelės negali turėti pirminio rakto, kurio atžvilgiu yra maišomi duomenys. Anonimizuotų duomenų saugojimas kitoje lentelėje leidžia išsaugoti pradinius duomenis, kurių gali prireikti, pavyzdžiui, vykstant tyrimui. Laikinoje lentelėje saugoti šie duomenys: kliento ID (pirminis raktas), vardas, pavardė, el. paštas. Po duomenų sukeitimo el. paštai buvo priskirti kitiems žmonėms. Pavyzdžiui, Marry Smith el. paštas *Customer* lentelėje yra MARY.SMITH@sakilacustomer.org, o laikinoje lentelėje atlikus sukeitimą jai priskirtas el. paštas Eldon19@yahoo.com.

Testuojant stulpelio anonimizacijos funkcijas buvo vykdomas *Address* lentelės adresų ir pašto kodų apibendrinimas. Atlikus nurodytas anonimizacijas iš adresų buvo panaikinti namų numeriai, pašto kodų paskutiniai du skaičiai pakeisti žvaigždutėmis. Ši lentelė naudota ir vienos iš trigerių sukūrimo funkcijos testavimo scenarijuje. Jo metu naudojant trigerio sukūrimo funkciją sukurtas naujas trigeris ir funkcija, kuri apibendrina įterpiamus adresus į *Address* lentelė. Testuojant šį funkcionalumą įterpti 1000 naujų įrašų naudojant *Bogus* biblioteką. Patikrinus naujausius įrašus šioje lentelėje adresai buvo, pavyzdžiui, ***** Oberbrunner Roads arba *** Lonny Tunnel.

Atliekant *Staff* lentelės gimimo datos apibendrinimą iki gimimo metų lygio, amžiaus, amžiaus rėžių, algos apibendrinimą pridedant jai netvarkos ir iki rėžių buvo sukurta nauja laikinoji lentelė su darbuotojų ID, vardais, pavardėmis, gimimo datomis ir algomis. Gimimo data anonimizuota iki gimimo metų, tuomet, perkūrus lentelę iki amžiaus. Dar kartą perkūrus lentelę gimimo metai anonimizuoti iki amžiaus rėžių. Algos buvo anonimizuojamos pridedant joms netvarkos, o tuomet anonimizuotos iki rėžių.

Atliekant trigerio sukūrimo funkcijų testavimo scenarijų sukurti nauji trigeriai ir funkcijos *Staff* lentelei, kurios apibendrina (iki gimimo metų lygio) įterpiamas gimimo datas ir algoms prideda netvarkos. Testuojant šį funkcionalumą įterpti 500 naujų įrašų, jų gimimo datos buvo sausio

1, o algoms pridėta netvarkos.

4.2. Anonimizacijos poveikis pasirinkto modelio duomenims

Atlikus prototipo testavimą naudojant pasirinkto modelio duomenis galima išskirti anonimizacijos poveikį jiems. 1 lentelėje pateikiamas anonimizacijos metodų tinkamumas atributams. Visgi, anonimizacija nėra trivialus procesas ir nurodytas tinkamumas ne visada užtikrina duomenų saugumą. Atliekant vardo, pavardės, naudotojo vardo, el. pašto anonimizaciją galima taikyti sukeitimo, slopinimo ir iškraipymo metodus. Visgi, pritaikius slopinimą arba iškraipymą, duomenys taps nebenaudojami, todėl slopinimą reiktų taikyti tik tiems duomenims, kurie nėra laikomi svarbūs tyrimui. Pritaikius slopinimą kliento vardui ir pavardei, el. paštui ar naudotojo vardui, sumažinama tikimybė identifikuoti asmenį pagal tikslius identifikatorius. Tačiau, pavyzdžiui, atlikus naudotojų vardų ar el. paštų sukeitimą, tikimybė identifikuoti asmenį išliktų didelė, jei šie atributai turi savyje asmens vardą ir/ar pavardę.

1 lentelė. Metodų tinkamumas atributams

Atributas	Metodas			
	4.1.1 1 Apibendrinimas	4.1.1 2 Sukeitimas	4.1.1 3 Slopinimas	4.1.1 4 Iškraipymas
Vardas	Ne	Taip	Taip	Taip
Pavardė	Ne	Taip	Taip	Taip
Naudotojo vardas	Ne	Taip	Taip	Taip
El. paštas	Ne	Taip	Taip	Taip
Pašto kodas	Taip	Taip	Taip	Taip
Telefono numeris	Taip	Taip	Taip	Taip
Adresas	Taip	Taip	Taip	Taip
Gimimo data	Taip	Taip	Ne	Ne
Alga	Taip	Taip	Ne	Taip

Pašto kodą ir telefono numerį galima anonimizuoti naudojant apibendrinimą, sukeitimą, slopinimą ir iškraipymą. Anonimizuojant šiuos atributus reikia atsižvelgti, kad juos galima saugoti ir kaip skaitinius, ir kaip tekstinius duomenų tipus. Jei duomenys saugomi tekstiniu tipu, apibendrinant juose galima paskutinius skaitmenis pakeisti žvaigždutėmis. Jeigu duomenys saugomi kaip skaitiniai – juos galima padalinti iš skaičiaus, kuris yra 10 kartotinis, taip panaikinant paskutinius skaitmenis. Nebeturint tikslios reikšmės sumažėtų tikimybė identifikuoti asmenį. Atliekant sukeitimą reiktų atsižvelgti į tai, kad pašto kodas yra susijęs su adresu, todėl ir šis atributas turėtų būti anonimizuotas. Slopinimo ir iškraipymo atveju šie atributai nebeturėtų jokios vertės.

Adresą galima anonimizuoti apibendrinant, sukeičiant, slopinant ir iškraipant. Apibendrinimo atveju duomenys išsaugotų savo naudingumą, tačiau dideliame mieste vienoje gatvėje yra daug

gyventojų, o mažo miestelio ar kaimo atveju tai reikštų nedidelę žmonių grupę. Slopinant ir iškraipant duomenys prarastų savo naudą, o sukeičiant reiktų atsižvelgti į tai, kad adresas susijęs ir su kitais atributais.

Gimimo datą galima apibendrinti 3 būdais ir sukeisti. Visi 3 apibendrinimo būdai išsaugo duomenų naudingumą. Anonimizuojant gimimo metus iki gimimo metų sausio 1 dienos duomenys nepraranda savo vertės, tačiau dalis žmonių yra gimę sausio 1, todėl jų gimimo data anonimizuotame rinkinyje išlieka tokia pati kaip pradinė. Taip pat, kitaip nei anonimizuojant gimimo datą iki amžiaus ar amžiaus rėžių, šiuo atveju nereikia keisti stulpelio tipo. Anonimizavus gimimo datą minimais dviem būdais duomenys išlaikytų savo naudingumą, bet stulpelio tipo keitimas ne visada yra galimas. Sukeitus gimimo datas vietomis būtų išsaugoti agregacinių funkcijų tikslumą.

Algą galima apibendrinti 2 būdais, sukeisti ir iškraipyti. Iškraipymo atveju duomenys prarastų savo vertę, nes vietoj jų būtų sugeneruoti atsitiktinai duomenys. Atlikus sukeitimą, kaip ir gimimo datos atveju, būtų išsaugomas agregacijos funkcijų tikslumas. Algoms pridėjus netvarkos, duomenys išliktų naudingi, tačiau tuo atveju, kai algos atotrūkis tarp daugiausiai uždirbančio darbuotojo ir kitų darbuotojų yra didelis, galima būtų identifikuoti asmenį. Vietoj algų rodant jų rėžius, duomenys išsaugotų savo naudingumą, tačiau reiktų keisti stulpelio tipą.

2 lentelė. Pradinės ir anonimizuotos algos agregacinių funkcijų rezultatų palyginimas

	Vidurkis	Mažiausia reikšmė	Didžiausia reikšmė	Mediana	Moda
Alga	1244,59	500	4435	1243	548, 118 ir 1868
Anonimizuota alga (rėžio dydis 100)	1246,31	-	-	-	[900, 1000)
Anonimizuota alga (rėžio dydis 500)	1243,51	-	-	-	[500, 1000)
Anonimizuota alga (rėžio dydis 1000)	1152,69	-	-	-	[1000, 2000)
Anonimizuota alga (netvarkos lygis – 10)	1248,79	471	4074	1244	541, 960 ir 1756
Anonimizuota alga (netvarkos lygis – 50)	1236,38	276	3791	1132,5	989 ir 691
Anonimizuota alga (netvarkos lygis – 80)	1246,83	111	5978	112	627

Nors kai kuriems duomenų tipams, pavyzdžiui, adresui, nėra lengva apskaičiuoti anonimizacijos poveikį, skaičiams tai nėra sudėtinga padaryti, nes jiems galima pritaikyti agregavimo funkcijas. Jų rezultatai pateikiami 2 lentelėje. *Staff* lentelėje neanonimizuotų algų vidurkis yra 1244,59,

mažiausia alga – 500, didžiausia – 4435, mediana – 1243, o moda – 548, 1188 ir 1868 (po 5 įrašus). Anonimizacijos poveikiui palyginti algos suskirstytos į rėžius 3 kartus: kai rėžio dydis 100, 500 ir 1000. Rėžio reikšmių vidurkis skaičiuotas sudėjus apatinį ir viršutinį rėžį ir sumą padalinus iš 2. Apskaičiuoti vidurkiai atitinkamai yra 1246,31, 1243,51 ir 1152,69. Moda atitinkamai [900, 1000) (82 įrašai), [500, 1000) (353 įrašai), [1000, 2000) (647 įrašai). Kai rėžio dydis 100 ir 500, vidurkiai išliko artimi pradiniam (pokytis atitinkamai 0.138% ir 0.086%), o kai rėžio dydis 1000, vidurkio pokytis didesnis – 7.383%. Mažesnio rėžio dydžio pasirinkimas leidžia užtikrinti didesnį anonimizuotų duomenų tikslumą. Visais trim atvejais tik du rėžiai stulpelyje pasitaikydavo vieną kartą.

Lygintas ir netvarkos pridėjimo poveikis. Pritaikyti 3 netvarkos lygiai – 10, 50 ir 80. Pirmuoju atveju vidurkis – 1248,791 (0.33% pokytis), mažiausia alga – 471, didžiausia – 4074, mediana – 1244, moda – 541, 960 ir 1756 (po 4 įrašus). Antruoju atveju vidurkis – 1236,38 (0.659% pokytis), mažiausia alga – 276, didžiausia – 3791, mediana – 1132,5, moda – 989 ir 691 (po 5 įrašus). Trečiuoju atveju vidurkis – 1246,83 (0.18% pokytis), mažiausia alga – 111, didžiausia – 5978, mediana – 112, moda – 627 (5 įrašai). Taigi, pokytis, pridėjus netvarkos, išlieka mažas, tačiau reikia atsižvelgti į tai, kad skaičiuojant anonimizuotą reikšmę yra dauginama iš atsitiktinio skaičiaus, todėl rezultatai, pritaikius šį metodą su tokiais pačiais parametrais, kiekvieną kartą bus skirtingi.

Datų atveju lyginti skirtumai tarp amžiaus ir amžiaus rėžių. Pirmuoju atveju vidutinis amžius – 28.86 metai, mažiausias – 18, didžiausias – 40, mediana – 29, moda – 29 (60 įrašų). Antruoju atveju vidutinis amžius – 29.24, moda – [20,30) (467 įrašai). Skaičiuojant vidurkį kiekvieno įrašo apatinio ir viršutinio rėžio suma padalinta iš 2.

4.3. Tyrimo rezultatai

Atlikto tyrimo metu buvo suprojektuotas, suprogramuotas ir ištestuotas duomenų anonimizacijos prototipas, pasirinktas duomenų modelis buvo užpildytas sintetiniais duomenimis naudojant *Bogus* biblioteką, anonimizuoti duomenys lyginti su pradiniais. Tyrimo rezultatai rodo, kad atributus galima anonimizuoti skirtingais būdais (1 lentelė). Tačiau nurodytas tinkamumas atributams ne visada užtikrina duomenų saugumą. Kiekvienas atvejis yra unikalus, nes gatvės pavadinimas dideliame mieste gali reikšti tūkstančius žmonių, o mažo miestelio atveju – kelių žmonių grupę. Anonimizavus algą pridėjus jai netvarkos išlieka tikimybė, kad bus galima atpažinti daugiausiai uždirbanti asmenį. Duomenų sukeitimas leidžia išsaugoti duomenų tikslumą naudojant agregacijos funkcijas, tačiau tinka ne visiems duomenų tipams ir išlieka tikimybė, kad po sukeitimo duomenys sutaps su pradiniais. Anonimizavus duomenis naudojant slopinimą jie tampa nebenaudojami,

bet esant poreikiui, pavyzdžiui, vykstant tyrimui, juos gali būti labai sudėtinga atstatyti į pradinius. Norint išsaugoti pradinius duomenis galima naudoti veidrodines duomenų kopijas, o prieigą prie neanonimizuotų duomenų suteikti tik įgaliotiems asmenims. Todėl anonimizuojant duomenis svarbu pasirinkti tinkamus anonimizacijos metodus atsižvelgiant į jų privalumus ir trūkumus.

Sukurtas duomenų anonimizacijos sprendimas tinka paprastiems duomenų tipams: tekstiniams, skaitiniams ar datoms. Tačiau asmeniniai duomenys gali būti saugomi ir sudėtingesniais duomenų tipais, pavyzdžiui, JSON ar XML, vartotojai gali sukurti savo sudėtinius tipus. Tokiais atvejais sukurtas anonimizacijos funkcijas reikėtų pritaikyti šiems duomenų tipams.

5. Rekomendacijos

Sukurtas duomenų anonimizacijos prototipas skirtas *PostgreSQL* RDBVS (reliacinė duomenų bazių valdymo sistema), tačiau jo funkcionalumą galima būtų pritaikyti ir kitoms RDBVS. Validavimui naudotos *information_schema* virtualiosios lentelės yra ir kitose RDBVS. Pavyzdžiui, norint atlikti tuos pačius validavimo veiksmus *Microsoft SQL Server*, *MySQL* arba *Oracle* RDBVS, reikalingus duomenis galima būtų gauti iš *information_schema.columns* ir *information_schema.key_column_usage* virtualiųjų lentelių. Taigi, skirtingose RDBVS užklausos išliktų tokios pačios.

BDAR suteikė vartotojams teisę būti užmirštiems. Tokiu atveju visi duomenys apie asmenį turėtų būti pašalinti arba anonimizuoti visose duomenų bazės lentelėse. Darant prielaidą, kad būtų anonimizuojami visi duomenys, o ne tik asmeniniai, naudojantis *information_schema* galima būtų rasti visus duomenis, kurie yra susiję su asmeniu. Tam reiktų naudoti *information_schema.table_constraint*, *information_schema.key_column_usage* ir*information_schema.constraint_column_usage*. Jungiant šias virtualiąsias lenteles galima rasti visus lentelių ryšius. Naudojantis išoriniu raktu būtų galima pašalinti duomenis ne tik iš pagrindinės lentelės, kurioje saugoma dalis duomenų apie vartotoją, bet ir iš kitų lentelių. Tačiau reikia atsižvelgti, kad kai kuriais atvejais tą patį išorinį raktą gali turėti daugiau nei vienas lentelės įrašas.

Taip pat, duomenys turi būti pašalinami, kai jie tampa nebenaudojami. Tam galima naudoti užduočių planuotojus (angl. job scheduler). Naudojantis jais galima nustatytu laiku atlikti tą pačią užklausą. Lentelėje turint atributą, kuris saugo paskutinio įrašo atnaujinimo dieną ir laiką, galima nustatytu laiku patikrinti, ar duomenys vis dar naudojami. Tai suteiktų galimybę, pavyzdžiui, kiekvienų metų sausio 1 dieną, iš duomenų bazės pašalinti jau nebenaudojamus klientų duomenis.

Tyrimo metu pasiūlyti būdai anonimizuoti skaitines reikšmes. Šie būdai tinka sveikiesiems skaičiams. Tačiau kartais asmeniniais duomenimis gali būti ir duomenys, kurie yra saugomi kaip slankaus kablelio skaičiai. Kai kurie metodai tinkantys sveikiesiems skaičiams nebūtų tinkami slankaus kablelio skaičiams. Todėl anonimizuojant slankaus kablelio skaičius reiktų naujų sprendimų.

Rezultatai ir išvados

Šio tyrimo metu nagrinėtos BDAR taisyklės ir duomenų saugojimo principai. Nagrinėtas BDAR taisyklių poveikis verslui ir technologijoms, egzistuojantys duomenų anonimizacijos sprendimai. Sukurtas ir ištestuotas duomenų anonimizacijos prototipas, pasirinktas modelis užpildytas sintetiniais duomenimis. Tirtas anonimizacijos poveikis modelio duomenims. Darbo rezultatai:

1. Apžvelgtas asmeninių duomenų apibrėžimas ir BDAR taisyklės. Apžvelgti pagrindiniai BDAR duomenų apsaugos principai, vartotojų teisės ir įmonių pareigos. Nagrinėta BDAR įtaka verslui ir technologijoms. Visos įmonės, kurios renka ir apdoroja ES piliečių duomenis, turi vadovautis BDAR nepriklausomai nuo jų saugojimo vietos. Naujos duomenų apsaugos taisyklės skatino įmones įdėti daugiau pastangų auginant kompetencijas kibernetinio saugumo srityje, mažinant turimų duomenų kiekį, atskiriant duomenis, vystant dirbtinio intelekto bei blokų grandinių technologijas. Įmonės norėdamos įgauti vartotojų pasitikėjimą turi tobulinti savo kibernetinį saugumą. Įvedus BDAR išaugo prašymų sutikti su privatumo politika kiekis tinklalapiuose.
2. Apžvelgta egzistuojanti literatūra, susijusi su duomenų anonimizacija. Aprašytas atributų skirstymas į grupes, anonimizacijos proceso etapai, anonimizacijos metodai. Nagrinėti *k*–anonimiškumo, *l*–įvairovės, *t*–artumo anonimizacijos modeliai, anonimizuotų duomenų atakos. Nagrinėti egzistuojantys duomenų anonimizacijos algoritmai, jų poveikis duomenims.
3. Tyrimo metu sukurtas ir ištestuotas duomenų anonimizacijos prototipas, skirtas *PostgreSQL* reliacinei duomenų bazių valdymo sistemai. Sukurtame prototipe pagrindiniai duomenų anonimizacijos metodai pritaikyti įprastiesiems duomenų tipams: tekstiniams, skaitiniams ir datoms. Prototipo funkcijos suskirstytos į 4 grupes (vieno įrašo anonimizacijos, stulpelio anonimizacijos, validavimo ir trigerių sukūrimo). Testavimo metu pasirinktas duomenų modelis, išskirti asmeninių duomenų atributai jame. Naudojant išorinę biblioteką sukurta programa skirta duomenų modelio užpildymui sintetiniais duomenimis, duomenų įterpimui naudota *Entity Framework Core* biblioteka.
4. Ištirtas anonimizacijos poveikis pasirinkto modelio duomenims. Išskirtas anonimizacijos metodų tinkamumas pasirinkto modelio atributams. Nagrinėtas anonimizacijos poveikis visiems pasirinkto modelio duomenims, kurie yra laikomi asmeniniais. Tirtas agregacinių funkcijų rezultatų pokytis anonimizuotiems skaitiniams duomenims ir datoms.
5. Suformuluotos rekomendacijos.

Darbo išvados:

1. Tyrimai rodo, kad įsigaliojęs BDAR įmonėms kelia finansinius bei produktų vystymo iššū-

kius, o vartotojams sukuria didesnę savo duomenų kontrolę ir suteikia supratimą, kas juos naudoja.

2. Atlikus duomenų anonimizaciją išlieka galimybė identifikuoti asmenis, o pažeidžiamumo lygis priklauso nuo išlikusio duomenų naudingumo. Anonimizuojant duomenų modelį svarbu pasirinkti jam tinkamiausius metodus, nes pasirinktas anonimizacijos metodas ne visada užtikrina duomenų saugumą.
3. Nors kai kuriems duomenų tipams sunku nustatyti anonmizacijos poveikį, tyrimo metu atlikto pradinių ir anonimizuotų duomenų palyginimo rezultatai rodo, kad skaitinių reikšmių pakeitimas į rėžius ar netvarkos pridėjimas skaičiams nepadaro didelės įtakos duomenų naudingumui.
4. Kai kurie duomenys turi būti anonimizuojami veidrodinėse duomenų bazėse, kad esant poreikiui, pradiniai duomenys būtų išsaugomi ir atsekami.

Šaltiniai

[18] Regulation (eu) 2016/679 (general data protection regulation), European Commission, 2018-05-25. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 (tikrinta 2019-10-22).

[ALM11] Georgia Albuquerque, Thomas Lowe ir Marcus Magnor. Synthetic generation of high-dimensional datasets. *IEEE transactions on visualization and computer graphics*, 17(12):2317–2324, 2011.

[APB18] Sabina-Daniela Axinte, Gabriel Petrica ir Ioan Bacivarov. Gdpr impact on company management and processed data. *Quality-Access to Success*, 19(165), 2018.

[APM+16] Vanessa Ayala-Rivera, A Omar Portillo-Dominguez, Liam Murphy ir Christina Thorpe. Cocoa: a synthetic data generator for testing anonymization techniques. *International Conference on Privacy in Statistical Databases*, p. 163–177. Springer, 2016.

[Bre19] Paul Breitbarth. The impact of gdpr one year on. *Network Security*, 2019(7):11–13, 2019.

[CEK16] Margareta Ciglic, Johann Eder ir Christian Koncilia. Anonymization of data sets with null values. *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV*, p. 193–220. Springer, 2016.

[Cha] Brian Chavez. Bogus for .net: c#, f#, and vb.net. https://github.com/bchavez/Bogus/. Accessed: 2020-05-11.

[DS16] Josep Domingo-Ferrer ir Jordi Soria-Comas. Anonymization in the time of big data. *International Conference on Privacy in Statistical Databases*, p. 57–68. Springer, 2016.

[DUL+18] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub ir Thorsten Holz. We value your privacy... now take some cookies: measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*, 2018.

[EAZ+18] Can Eyupoglu, Muhammed Ali Aydin, Abdul Halim Zaim ir Ahmet Sertbas. An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, 20(5):373, 2018.

[GJS19] Samuel Goldberg, Garrett Johnson ir Scott Shriver. Regulating privacy online: the early impact of the gdpr on european web traffic & e-commerce outcomes. *Available at SSRN 3421731*, 2019.

[GKK+07] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis ir Nikos Mamoulis. Fast data anonymization with low information loss. *Proceedings of the 33rd international conference on Very large data bases*, p. 758–769, 2007.

[GMV+18] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi ir Meiko Jensen. Privacy issues and data protection in big data: a case study analysis under gdpr. *2018 IEEE International Conference on Big Data (Big Data)*, p. 5027–5033. IEEE, 2018.

[Gre18] Samuel Greengard. Weighing the impact of GDPR. *Communications of the ACM*, 61(11):16–18, 2018.

[Gro] The PostgreSQL Global Development Group. Postgresql: documentation: 12: chapter 36. the information schema. `https://www.postgresql.org/docs/current/information-schema.html`. Accessed: 2020-05-16.

[Hsu18] Jeremy Hsu. What you need to know about europe's data privacy rules [resources_at work]. *IEEE Spectrum*, 55(5):21–21, 2018.

[JLM+15] Shouling Ji, Weiqing Li, Prateek Mittal, Xin Hu ir Raheem Beyah. Secgraph: a uniform and open-source evaluation system for graph data anonymization and de-anonymization. *24th {USENIX} Security Symposium ({USENIX} Security 15)*, p. 303–318, 2015.

[KV17] Tanashri Karle ir Deepali Vora. Privacy preservation in big data using anonymization techniques. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, p. 340–343. IEEE, 2017.

[LYH19] He Li, Lu Yu ir Wu He. The impact of gdpr on global technology development, 2019.

[LPC+18] Wei Liu, Mengli Pei, Congcong Cheng, Wei She ir Chase Q Wu. An improved data anonymization algorithm for incomplete medical dataset publishing. *The International Conference on Healthcare Science and Engineering*, p. 115–128. Springer, 2018.

[MBR+19] Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim ir Ramona Ramli. A comparative study of data anonymization techniques. *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, p. 306–309. IEEE, 2019.

[MDK19] Maja Gligora Marković, Sandra Debeljak ir Nikola Kadoić. Preparing students for the era of the general data protection regulation (gdpr). *TEM Journal*, 8(1):150–156, 2019.

[NK17] J Jesu Vedha Nayahi ir V Kavitha. Privacy and utility preserving data clustering for data anonymization and distribution on hadoop. *Future Generation Computer Systems*, 74:393–408, 2017.

[NKU18] Deepak Narula, Pardeep Kumar ir Shuchita Upadhyaya. Privacy preservation using various anonymity models. *Cyber Security*, p. 119–130. Springer, 2018.

[PAP18] Eugenia Politou, Efthimios Alepis ir Constantinos Patsakis. Forgetting personal data and revoking consent under the gdpr: challenges and proposed solutions. *Journal of Cybersecurity*, 4(1), 2018.

[SKP+18] Junwoo Seo, Kyoungmin Kim, Mookyu Park, Moosung Park ir Kyungho Lee. An analysis of economic impact on iot industry under gdpr. *Mobile Information Systems*, 2018, 2018.

[SOT+17] Yuichi Sei, Hiroshi Okumura, Takao Takenouchi ir Akihiko Ohsuga. Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness. *IEEE transactions on dependable and secure computing*, 2017.

[Tan16] Colin Tankard. What the gdpr means for businesses. *Network Security*, 2016(6):5–8, 2016.

[VH19] W Gregory Voss ir Kimberly A Houser. Personal data and the gdpr: providing a competitive advantage for us companies. *American Business Law Journal*, 56(2):287–344, 2019.

[Web] PostgreSQL Tutorial Website. Postgresql sample database. `http://www.postgresqltutorial.com/postgresql-sample-database/`. Accessed: 2019-12-27.

[Zar16] Tal Z Zarsky. Incompatible: the gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995, 2016.