MTAT.07.007 Krüptograafia kraadiõppeseminar

Delikaatsete andmete varjestamine

Uuno Puus
Cybernetica AS
uuno@cyber.ee

Probleem delikaatsete isikuandmetega

- Delikaatsete isikuandmete töötlemisel tuleb säilitada andmeallika privaatsus

- Eelkõige tervishoiu- ja sotsiaalvaldkonnas uurimistöö läbi viimiseks on vajalik töödelda ka delikaatseid isikuandmeid

Delikaatsete isikuandmete kasutus on reguleeritud

- Delikaatsete isikuandmete kasutamine/töötlemine on reguleeritud Isikuandmete kaitse seadusega. Järelvalvet seaduse täitmise üle teostab Andmekaitse Inspektsioon

- Isikuandmete kaitse seadusest on valminud uus versioon

Uurimistöö läbiviijad pole suutnud olukorraga kohaneda

- Senini on nimtetatud valdkondade teadlased kasutanud oma töös delikaatseid isikuandmeid ilma delikaatsete andmete kaitsmisele tähelepanu pööramata
- Kui sel viisil enam uurimistööd polnud õimalik läbi viia, siis süüdistati inspektsiooni uurimistöö takistamises
- Sedalaadi protestid viisidki Isikuandmete kaitse seaduse uue versiooni välja töötamiseni

Mingid lahendused on tegelikult olemas

- Jerome P. Reiteri artiklis ”Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study” on kirjeldatud võimalusi delikaatsete isikuandmete varjestamiseks – so delikaatsete isikuandmete statistiliseks analüüsiks ilma andmeallika privaatsust rikkumata

- Delikaatsete andmete varjestamiseks kasutatakse põhimõtteliselt kahte võimalust

  ⋆ Müra lisamine andmetele sel viisil, et neid analüüsides poleks võimalik teha järeldusi konkreetsete isikute kohta

  ⋆ Esialgsete andmete modifitseerimine lisades neile nn sünteetilisi andmeid

Esialgsetele andmetele müra lisamine

- Näiteks tunnus vanus kodeeritakse vanusevahemikena – 0-15 a, 15-20 a jne., juhul kui see pole võimalik (näiteks binaarse tunnuse puhul), muudetakse valeks mingi osa (näiteks 10%) andmetest

  ⋆ Sel juhul jäävad andmed veel statistiliselt usaldusväärseks, kuid isikute täpsusga pole enam võimalik öelda, millisel objektil antud tunnus esines, millisel mitte.

  ⋆ Müra kasutamist on kirjeldatud Wayne A Fulleri poolt 1993. aastal ilmunud artiklis ”Masking Procedures for Microdata Disclosure Limitation”

- Müra lisamine pole hea lahendus, kuna komplitseerib andmete hilisemat töötlemist

Teine võimalus – nn sünteetilised andmed

- Sünteetiliste andmete lisamine (konfidentsiaalsuse säilitamise eesmärgil) koosneb kahest etapist
  - ⋆ Vaid osa andmeid (andmetabeli veerge) võetakse lähteandmestikust. Ülejäänud sünteesitakse kasutades olemasolevaid andmeid statistilisi meetodeid kasutades
  - ⋆ Eelkirjeldatud viisil saadud täielikust tabelist antakse uurijatele analüüsiks kasutada juhuvalimid

Andmeallika privaatsus on tagatud, sest

- Uurija saab analüüsiks juhuvalimi st ta kasutab vaid juhuvalimisse sattunud objektide andmeid
- Kui uurija leiab andmestikust tuttava andmed siis ta ei saa olla kindel, kas need olid algses andmestikus olemas või sünteesiti hiljem
- Eelpool toodu välistab andmeallikate identifitseerimise. Tuletuslikul visil (predictive disclosure) konfidentsiaalsuse kao tõenäosus/oht jääb siiski püsima

Vaatleme artiklis toodud näidet

- Artiklis on kirjeldatud Ameerika elanikkonna uuringu näidet. Elanikkonna andmete analüüsil püstitati järgmised ülesanded:
  - ⋆ Hinnata sünteetiliste andmete põhjal tehtud järelduste valiidsust
  - ⋆ Hinnata konfidentsiaalsuse säilitamise taset, mida võimaldab sünteetiliste andmete kasutamine
  - ⋆ Illustreerida ja uurida sünteetiliste andmete loomise protsessi

Käsitletav Ameerika rahvastiku-uuring aastal 2000

- Uuringus osales 133170 inimest 51016-st leibkonnast
- Kõigi uuringus osalenud isikute kohta koguti järgmised andmed
  - ⋆ sugu, rass, perekonnaseis, haridus, vanus, lastetoetuse suurus, elatusraha suurus, alimendid, eluaseme kulud, sissetulek leibkonna kohta
- Kogu andmete hulgast valiti juhuslikult 10 000 kirjet, millest omakorda valiti juhuslikult 500 erinevat valimit, igaühes 100 kirjet.

Sünteetiliste andmete genereerimine

- Vanus, sugu, rass lülitati sünteesandmetesse sellisena nagu nad andmestikus olid

- Ülejäänud genereeriti järgmises järjekorras

  ⋆ haridus, perekonnaseis, alimendid, lastetoetus, toimetuleku toetus, leibkonna sissetulek, eluaseme kulud

  ⋆ Põhimõtteliselt võib kasutada ka teistsugust järjekorda, eelpool toodu valiti lähtuvalt analüüsitava andmestiku eripärast

- Vaatleme lähemalt kuidas genereeriti haridustase

Haridustaseme sünteesimine

- Haridustase sõltub eelkõige vanusest st eri vanuste kohta on erinevad valemid

- Kõigepealt arvutati välja lähendid perekonnapea haridustaseme kohta

  ⋆ Mudelid sisaldavad tunnuseid sugu, rass

  ⋆ Kui $17 < vanus < 25$, siis $vanus + vanus^2$

  ⋆ Kui $vanus > 24$, siis $(54 < vanus < 65) + (vanus > 64)$

- Seejärel sünteesiti haridustase kõigi uuringus osalenud objektide kohta

Tabel

Objekt	vanus	sugu	rass	haridus	perekonnaseis
Objekt 1				——-	——-
Objekt 2				——-	——-
Objekt 3				——-	——-
. . .	. . .	. . .	. . .	. . .	. . .
Objekt n				——-	——-

Kokkuvõte

- See uuring ei selgitanud välja sünteetiliste andmete genereerimise eeliseid võrreldes teiste konfidentsiaalsuse tagamise meetmetega
- Rakendades samadele andmetele erinevaid meetodeid ja võrreldes nende tulemuslikkust sünteetiliste andmete genereerimisega saab ehk teha täpsemaid järeldusi
- Mitte vähem tähtis pole ka kasutatavate meetodite nn kasutusmugavus nii andmete sünteesijate kui ka nende kasutajate jaoks

Kuidas selles artiklis kirjeldatud kogemust Eestis rakendada

- Praegu käib kodeerimiskeskuse loomine. Selle ülesandeks on just konfidentsiaalsuse tagamine delikaatsete isikuandmete kasutamisel teadusuuringutes

- Ettepanek on nimetatud kodeerimiskeskus asendada näiteks statistikakeskusega, mis tegeleks eelpool kirjeldatud delikaatsete isikuandmete varjestamisega ning uurijatele vajalike (sünteetiliste) andmete ette valmistamisega

- See lahendus ei aita andmete dekodeerimisel, kui seda peaks tarvis minema

Küsimused, märkused, kommentaarid.

Tänan tähelepanu eest!