Tallinna Tehnikaülikool
Infotehnoloogia teaduskond
Arvutiteaduse instituut

Sotsiaalmeediast isikuandmete masskogumise meetodid Facebooki näitel

magistritöö

Üliõpilane: Jens Kaspar Mikli
Üliõpilaskood: 132381IAPM
Juhendaja: Tanel Tammet

Tallinn
2016

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

_____ _____

(*kuupäev*) (*allkiri*)

Annotatsioon

Töö sisu on sotsiaalmeediat kasutavate automatiseeritud isikuinfo kogumise meetodite uurimine. Erilist tähelepanu on pööratud ekraanikraapimise tehnoloogiale ning selle tehnoloogia rakendamise tulemuste talletamisele ja analüüsimisele. Põhiküsimus on seejuures: kas ja kui suurel määral on võõrastel kontodel olevatele infohulkadele tavapäraste kasutajaõigustega isikul tehniliselt võimalik ligipääsu saavutada.

Töö jooksul arendati välja meetod, kuidas jälgida ja alla tõmmata suures koguses isikuinfot Eestis suurima osakaaluga sotsiaalvõrgustiku, Facebooki, näitel. Sealjuures tuvastades, mis viisil Facebooki lehekülg üritab info massilist allatõmbamist takistada ja kuidas neist vastumeetmetest on võimalik läbi pääseda. Samuti kirjeldame, kuidas need piirangud on aja jooksul, sealjuures ka antud töö läbiviimise jooksul, muutunud. Töö jooksul valminud võrgu-roomaja funktsionaalsust, tehnoloogiat ja valmistamist on kirjeldatud ning lähtekood on avalikustatud.

Töö analüüsi-osas demonstreeritakse näitlikult, milliseid andmeid sel viisil koguti ja mis on mõned selliste andmete rakendusviisid. Näidetena on toodud kasutajagruppide analüüs võtmeõnade kaudu ja stosiaalvõrgustikkudest graafi loomine. Analüüsi-osa on eeskätt mõeldud illustreerimaks võimalikke põhjalikumaid rakendusi.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 36 leheküljel, 4 peatükki, 6 joonist, 2 tabelit.

Abstract

The content of this work involves the study of automated data gathering methods that utilize social networking websites as a source with the primary focus being on data scraping technology. Data scraping was used to gather and analyse user information in order to answer the question of whether and how easily this could be achieved by a someone lacking special access privileges and what information such a person could theoretically obtain via analysis of such data.

During the course of the work a method of data scraping Estonia's most widely used social network Facebook was developed. Facebook uses a variety of means in order to prevent such activity and these are discussed in the work along with means of circumventing some of these measures and how these have changed during the course of the study. The development of the scraping software is discussed and its source code provided online.

Personal data was gathered for analysis using the former. In the final section the types of data acquired are discussed. Examples of possible uses of such data are include such as the analysis of social groups through keywords as well as the graphing of social networks.

Lühendite ja mõistete sõnastik

Ekraanilt kraapimine	*Data scraping* *tehnoloogia, mis kogub andmeid tavakasutajale loodud interneti lehelt*
Libakonto	*Sockpuppet* *fiktiivsele isikule loodud kasutajakonto*
OSINT	*Open source intelligence* *avalikustatud infoallikad*
Parser	*parser* *andmete leidmise ja tõlgendamisega tegelev tarkvara*
Phishing	*e. õngevõtmine* *isikuinfole pettuse kaudu ligipääsemise võtted*
Roomaja	*crawler* *automaatne veebisisu lehitseja ja info allatõmbaja*
SOCMINT	*Social media intelligence* sotsiaalmeedia infoallikad

Jooniste nimekiri

Tabelite nimekiri

Sisukord

Sissejuhatus

Sotsiaalmeedia on teatud tüüp veebiteenuseid ja -lehekülgi, mille sisu genereerivad kasutajad. Sellel on maailmas kasvav majanduslik tähtsus. Sotsiaalmeedia poolt pakutav ligipääs isiklikule infole loob palju võimalusi nii äritegevuseks, õiguskaitseks kui ka kuritarvitamiseks.

Sellisele infole ligipääs on aga tihtipeale piiratud. Seda kasutaja enda heaoluks või eesmärgiga kaitsta info ligipääsuõigusi selle nimel, et seda hiljem edasi müüa. Mõlemal juhul ei ole teenusepakkuja huvides teha kogu sotsiaalvõrgustikus paiknev kasutajainfo kõigile kättesaadavaks. Sõltuvalt sellest, mis turvameetmeid selle vastu rakendatakse, on võimalik neile andmetele siiski ligi pääseda, isegi teenusepakkuja teadmata.

Info isiklike andmete turvalisuse kohta ja kuidas näeb välja nende ründamine, on tihti salajane, kuna selle avalikustamine annaks võimaluse neid meetodeid väärkasutada. Samas on selline info tarvilik nii infoturbe taseme tõstmiseks, kui ka selle taseme määramiseks. Kuna sotsiaalvõrgustike teenusepakkujad muudavad pidevalt viisi, kuidas andmeid avalikustatakse ja nende kasutajate hulk ning info maht aina kasvab, on selle teema uuring ajakohane ja potentsiaalselt väärtuslik. Esimese peatükis „Ülevaade valdkonnast“ on laiemalt kirjeldatud teenusepakkujate seisukohti, sotsiaalmeedia rolli ja tehnoloogiat ning seda, kes sealt kogutavate andmete vastu huvi tunneb.

Töö peamine eesmärk on võõrastel kontodel olevat isiklikku infot massiliselt koguvate automaatsete skriptide uurimine, parandamine ja võrdlemine teiste sarnaste meetoditega. Seda selleks, et määrata, milliseid tööriistu isikliku info kogumisel kasutatakse. Sealjuures vastatakse ka küsimustele, kas ja kui suurele hulgale privaatsetele andmetele on isikul, kellel puuduvad erilised ligipääsu privileegid, võimalik ligi pääseda. Samuti, mis viisil neid automatiseerimismeetodeid rakendada. Täpsemad tööküsimused ja -eesmärgid on kirjeldatud alapealkirja all „Ülesande püstitus.“

Töö eesmärgi täitmiseks on uurimise alla võetud avalikult saadavad infokogumiseks loodud tööriistad, sealjuures ka Facebooki sotsiaalvõrgustikust infot koguv skript „fbstalker“ ja teised skriptid, mis otsivad inimese kohta andmeid, andmekaevandades sotsiaalvõrgustikke. Peatükis „Varasemad tööd“ on kirjeldatud sarnase uurimisteemaga tegelenud tööde tulemusi ja varem loodud analoogsete tööriistade tegevusprintsiipe.

Järgnevates peatükkides: „Töökäigu plaan,“ „Töökäigu kirjeldus“ ja „Arenduse käigus tekkinud probleemide analüüs,“ on kirjeldatud praktilist töö osa, kus toodeti toimiv infokogumisskript, mida rakendati töö jooksul isikuinfo kogumiseks Facebooki sotsiaalvõrgustikust. Infot, mida töö jooksul koguti, on kirjeldatud tulemuste analüüsis ning sellest tulenevad järeldused tööprotsessi kohta ning vastuseid eelnevatele küsimustele peatükis “Järeldused.“

1.1 Ülevaade valdkonnast

Eestis enim kasutatud sotsiaalvõrgustik Facebook (Forrester, 2014) keelab oma kasutuspoliisis automaatse infokogumise. Selle õiguslikuks läbiviimiseks nende lehestikul on ennem vajalik sõlmida firmaga eraldi leping, mille tingimuste juurde kuuluvad ranged piirangud selle kohta, kuidas infot tohib kasutada ja talletada (Automated Data Collection Terms, 2010).

Töö ülesandeks on selgitada, kui kergesti on võimalik saada ligi isiklikele andmetele eriliste ligipääsuõigusteta eraisikul. Eeskätt, kui selleks rakendades automaatseid skripte ja avalikke infoallikaid, nagu sotsiaalvõrgustikud. Samuti vastata küsimusele, kui kergesti saab selliselt leitud mass-andmetest teha analüüse ja järeldusi. Seda teemat on käsitletud ka „Social media intelligence“ ehk „socmint“ ja „open source intelligence“ nime all kus „open source“ viitab inimsete enda poolt avalikustatud infole.

Sotsiaalmeedia on võimas inimtegevuse organiseerimise ja meelsuse mõjutamise tööriist. See on kiire ja laiahaardeline ning on võimeline inimesi efektiivselt koordineerima. Analüüsitavad andmed sotsiaalmeedias võib jaotada sisu, seoste ja metaandmete gruppidesse. Sisu (content) on avalikustatavad andmed, seoste (relationship) andmed näitavad sisuandmete seoseid teiste sisuandmete ja isikutega ning metaandmed on meetrika andmete avalikustamise konteksti kohta. Kogutud andmeid saab kasutada muuhulgas suhtumise hindamiseks (sentiment analysis), inimeste käitumise kohta info soetamiseks, või teiste infokogumise meetodite alusena.

Praktikas ületab suur andmehulk siinjuures analüüsi võimeid. Sealjuures võib oodata võrreldes traditsionaalsete andmeallikatega ka rohkem pettust, kuna rangeid kriteeriume avalikustamiseks ei ole. Väga tähtis on teada, mis infot otsitakse, et tööd saaks mõistlikult piirata. Kuna andmemaht on niivõrd suur, on tähtis valida kus ja millal monitoorida. Näiteks Youtube.com lehele laetakse üles 3.5 Petabaiti videot päevas. Sellise andmete hulga töötlemiseks on tarvis luua erialgoritme ja rakendada arvutiklastreid(Forrester, 2012).

Tegemist on andmekaevandamise valdkonnaga tihedalt seotud tööga, kuid infoturbe jaoks ei ole tähtsad vaid info trendid, vaid ka see, kui hästi suudab info vastata kriitilistele küsimustele. Selle tarvis arendatakse algoritme ja automatiseeritud tööriistu. Järeldada võib, et avalik meedia on laialdane uurimisvaldkond.

NATO RTO (Research and Technology Orginization) on kokku pannud uurimismeeskonna mis peab selgeks tegema socmint-i (social media intelligence) rolli tulevikus (Forrester, 2012, 2014). Sealjuures suurimaks katsumuseks loetakse andmete suurt mahtu, mis kaugelt ületab inimjõul analüüsimise võimeid. Seega keskendub NATO töö sellele, kuidas monitoorida, analüüsida ja ennetada sotsiaalmeedia käitumist. Samuti kindlaks määrata, milline on parim viis sellist infot kasutada ja millised on sealjuures kasutatavad metodoloogiad ning riskid võrreldes traditsionaalsete avalike infoallikatega. Võrdlev tabel (tabel 1) näitab kokkuvõtlikult erinevusi sotsiaalmeedia (socmint) ja avaliku (osint) meedia ressursside vahel.

OSINT	**SOCMINT**
Akadeemiline uuring, raamatud, entsüklopeediad, ametlikud dokumendid, pildid, ajakirjad, ülekantav meedia, kaardid, ajalehed, raadio	Blogid, micro-blogid, Interneti foorumid, kasutajate loodud FAQd, jututoad, podcastid, online mängud, lipikud (tags), hinnangud(ratings), kommentaarid, sotsiaal võrgustike lehed, videod, wikid, otsingumootorid, sotsiaalsed järjehoidjad
Toimetatud	Toimetamata
Professionaalsed autorid	Ka tundmatud autorid
Kirjakeel korrektne ja formaalne	Tihti ebakorrektne kirjakeel
Hästi kataloogiseeritud koos metaandmetega.	Nii kuidas leitud (ka folksonoomia, lipikud)
Kerge ligipääas	Tuleb välja otsida
Mõnedele ligipääs maksab.	Tavaliselt ligipääs ei maksa midagi, mõnedel on API piirangud ligipääsule
Hästi defineeritud akronüümid	Tundmatuid akronüüme– netikeel
Heli ja video on hea kvaliteediga	Heli ja video kvaliteet varieerub

Tabel 1: *võrdlus sotsiaalmeedia ja teiste avalike infoallikate vahel (Forrester, 2012).*

Olemasolevad tarkvaratööriistad keskenduvad ennekõike reklaamiteenustele info kogumisele. Tööriistad on nt. Google analytics, Tweetreach, Youtube Insight jne. Need on enamasti maksulised või pakuvad piiratuid teenuseid (Dyer, P., 2013).

Sealjures pakutavate teenuste hulka kuuluvad: otsingupäringutes isiku või toote ilmumise määramine, arvamuse, mõjukuse, levi, trendi analüüs, ja bränd-i võrdlus. Neil puudub tihti küberkaitse jaoks tarvilik tellija anonüümsus ja sügavam analüüsivõime. NATO raport jõudis tulemusele, et selliste rakenduste arendamisega on kiire, kuna sotsiaalmeedia on pidevalt arenev ja kõrge potentsiaaliga ressurss (Forrester, 2012).

Teisest küljest toimub phishing, s.t. isikuandmete vargus, mitmel sarnasel viisil. Olemas on email ja interneti lehed ja teenused, mis on loodud taaskasutatavate paroolide ja kasutajanimede varguseks. Kõik need tegevused võivad mingil määral sotsiaalvõrgustike kaudu leitavate isikuandmete kaudu ka tulemuslikumad olla. Suur roll on „sock puppet“ ehk libakontodel, mille alt kogutakse informatsiooni sihtmärgi nimel(Sullivan, 2014). Isikuinfo pettuse teel soetamise jaoks rakendatakse ka emaili, veebilehti ja wifit, et kasutaja oma salastatud infot ise avaldaks. Sealjuures on pettused tihti maskeeritud mingi sotsiaalmeedia teenusena. Samas toimub isikuandmete kogumine ka politseitöö käigus ja on infoturbe loomulik osa.

1.2 Ülesande püstitus

Töö ülesandeks on selgitada, kui kergesti ja kui suures koguses on võimalik saada ligi isiklikele andmetele, rakendades selleks automaatseid skripte ja sotsiaalvõrgustike kontosid. Sealjuures määrata, kas on olemas teoreetiline oht, et isik, rakendades automatiseerituid meetodeid, privaatsetele andmetele ligi pääseb. Samuti mis viisil ta neid meetodeid rakendada võiks. Selle teema uurimiseks kasutan näitena Facebooki sotsiaalvõrgustikku, mis on Eestis suurima kasutajate hulgaga (Forrester, 2014). Samas on avalikest allikatest andmekaevandus laiem teema ning uurimustöö tulemusena võib tekkida suurem arusaamine selle läbiviimise kohta.

Töö sisaldab lühidalt ülevaadet „Data scraping“ ehk andmekraapimise st. andmete kogumise interneti lehekülgede lähtekoodi parsimise kaudu ja „Sockpuppet“ ehk libakontode ja nende rakendamise kohta andmekraapimise nimel ja andmete kogumise kohta sotsiaalvõrkudest.

Töö ülesanneteks on ka isiklikku infot kuvava skripti uurimine ja parandamine ning võrdlemine teiste sarnaste meetoditega, samuti teooria ülevaade ning statistika koostamine infokogumise kohta. Selgitamist vajavad küsimused isiklike andmete kaevandamise kohta on: millised tööriistad on antud töö eesmärki silmas pidades rakendatavad ja millised on nende põhiomadused? Samuti: kuidas erinevaid meetodeid omavahel võrrelda, missugustele andmetele on lihtne ja millistele keeruline ligi pääseda, kuidas oleks võimalik nende meetodite eest kaitsta?

Vaheküsimuseks oleks, kui palju andmeid inimesed on nõus avalikustama ja kui palju tööd nõuab nende andmeteni jõudmine? Selle jaoks võrdlen eelpool mainitud skripte avalike API-dega, mis andmeid sotsaalvõrgustikest tõmbavad. Statistikas kuvatakse, mis andmed olid üle valimi avalikud ja kuidas neile skript ligi sai. Uurimise läbiviimise jooksul valmib ka variant skriptist, mis seda infot talletab. Lisaküsimus on ka korrelatsiooni määramine: kas infot, mis võrgustikust kogutakse, on teoreetiliselt võimalik isikuga siduda ka väljaspool võrgustikku? Kui lihtsalt ja mida selliselt hangitud andmetest on võimalik tuvastada ja üldistada?

1.3 Varasemad tööd

Edukas automaatse tarkvara näide oleks Reynolds jt. töö mõjukate Twitteri kasutajate määramisel statistika alusel (Reynolds et al., 2010) kus suhete ja korduvpostituste jälgimise kaudu määrati kõige mõjukamad kasutajad. Sarnane Facebooki roomaja ja analüüsija on Euroopa Liidu poolt finantseeritud uurimise projekt nimega CAPER, mis toimib uuritavate skriptidega sarnasel põhimõttel ja on arendatud kuritegevuse jälgimiseks euroopa õiguskaitse instituutide poolt (Aliprandi et al, 2014). Samuti Facebooki näitel loodud Bonneau, Anderson ja Danezis 2009. aasta töö näitas, et andmekaevanduse vastu polnud Facebook kaitstud ja on tuvastatud mitmeid meetodeid isiklike ja võrgustike kohta andmete kogumiseks. Osad millest toimivad isegi kasutaja privaatsussätetest sõltumata. Vaja on ainult mõnede inimeste kontodele ligipääsu, et võrgustikku avalikustada. Samas enamus inimesi ei mõista või ei huvita neid lehekülje privaatsussätted. See uurimistulemus on nüüd rohkem kui kuus aastat vana ja on ajakohane vaadata, kuidas on olukord tänapäeval.

Lisaks on mõningate sotsiaalvõrkude poolt pakutavate piltide üleslaadimisteenuste (twitter ja flickr) alusel võimalik kaardistada inimeste asukohta, kui seda infot eraldi piltidelt ei eemaldata (inglis keeles kutsutakse neid andmeid „geolocation data“-ks). Seda meetodit kasutab ka skript cree.py (Kakavas, 2011) ja seda on rakendatud ka Compton, Jurgens ja Allen’i töös (Compton, Jurgens & Allen, 2014). Nimelt on piltidele salvestatud andmete kaudu võimalik tuletada konto omaniku asukohta. Selle eest saab kasutaja end kaitsta hoolika privaatsussätete seadistamisega. Paljud teenusepakkujad teevad selle keeruliseks, sest isikuinfo müümine on sotsiaalvõrkkude sissetuleku allikas. Lisaks on interneti monitooringu tarkvara pidevalt arenemas.

Vastuseks ülesande püstituses esitatud küsimusele: „Millised tööriistad on antud töö eesmärki silmas pidades rakendatavad ja millised on nende põhiomadused?“ on meile teada kaks sellist tööriista: andmekogumisskriptid fbstalker ja cree.py. Töös uurimise jaoks valiti esimene, kuna cree.py on skript, mis kogub infot peamiselt teenuselt Twitter, mida Eestis, võrreldes Facebookiga, vähem kasutatakse. Samuti on fbstalker avalikult kättesaadava lähtekoodiga, mis tähendab, et seda saab rakendada ja edasi arendada ja võib eeldada, et andmed, mida õnnestub talletada, on kättesaadavad kõigile, kes on sellest huvitatud. Sama firma poolt, kui fbstalker, on arendatud ka geostalker skript, mis otsib asukoha andmete järgi inimesi ja asukohale lähedast wifit.

Lisaks võrguroomajatele kogutakse infot ka andmeanalüüsi kaudu. Kogudes infot sotsiaalvõrgu kasutajate suhete kohta, mis on tihti salastamata, on võimalik leida korrelatsioone erinevate võrkude kasutajate vahel. Selline de-anonüümiseerimine on väga efektiivne ja aitab siduda mitmeid kontosid sama isikuga. Näiteks fbstalker üritab kasutaja käitumise kaudu sõbranimekirja luua ka siis, kui puudub otsene ligipääs sellele infole. Kasutades suuremaid andmebaase kasutajasuhete kohta on võimalik leida korrelatsioone erinevate sotsiaalvõrkude vahel. Narayanan, A. & Shmatikov, V. töö (Narayanan & Shmatikov, 2009) võrdles kahe sotsiaalvõrgu Twitter ja Flickri kasutajate nimesid ja tuletas nende kaudu ülejäänud sotsiaalvõrgu kasutajate identiteete. Salvatrore jt. 2011 töös loodi samuti sotisaalvõrgu graaf, et selle analüüsi kaudu isikuandmeid tuletada. Wondracek jt. 2011 töös aga kombineeriti roomaja andmekogumine phishing võtetega et kasutaja identieeti tuvastada.

Neid erinevaid meetodeid koos rakendades on võimalik tuletada suurel hulgal infot inimese kohta, isegi, kui ta ei ole konkreetseid andmeid avalikustanud.

2. Töökäigu plaan

Tööülesande täitmiseks üritan leida viise, kuidas Facebooki lehestikus talletatud andmeid kätte saada. Sealjuures huvitab mind kõik, mis kasutaja oma lehele on lisanud, kuid eriti andmed, mis aitavad isikut tuvastada ja mida Facebook tavapärasele otsimootorile ei avalda. Töö jaoks piirdun Facebooki sotsiaalvõrgu uurimisega, kuna see on üks suurimaid maailmas ja Eestis enim kasutatud. Samuti võib eeldada, et infoturbe meetodid on uurimiseks piisavalt keerukad.

Info kogumist võib jagada passiivseks ja aktiivseks, kus phishing- tegevus ja lehekülje kasutajatega otseselt suheldes infole ligipääsu võitmine kuulub aktiivse tegevuse alla. Ilmselt on sellise tehnikaga võimalik pääseda ligi suuremale hulgale andmetele, kuid tõenäoliselt ei aita see leida turvalünki lehe disainis ning nõuab vähem tarkvara kui inimoskusi (siiski mitte eksklusiivselt, nt. võltsides kasutajaliidese välimust on võimalik luurata inimese tegevuse kallal kaudselt). Kuna töös huvitavad meid ennekõike automaatsed andmekogumise meetodid, keskendume passiivsetele infokogumise võtetele.

Info passiivselt kogumise jaoks on ennekõike kaks meetodit:

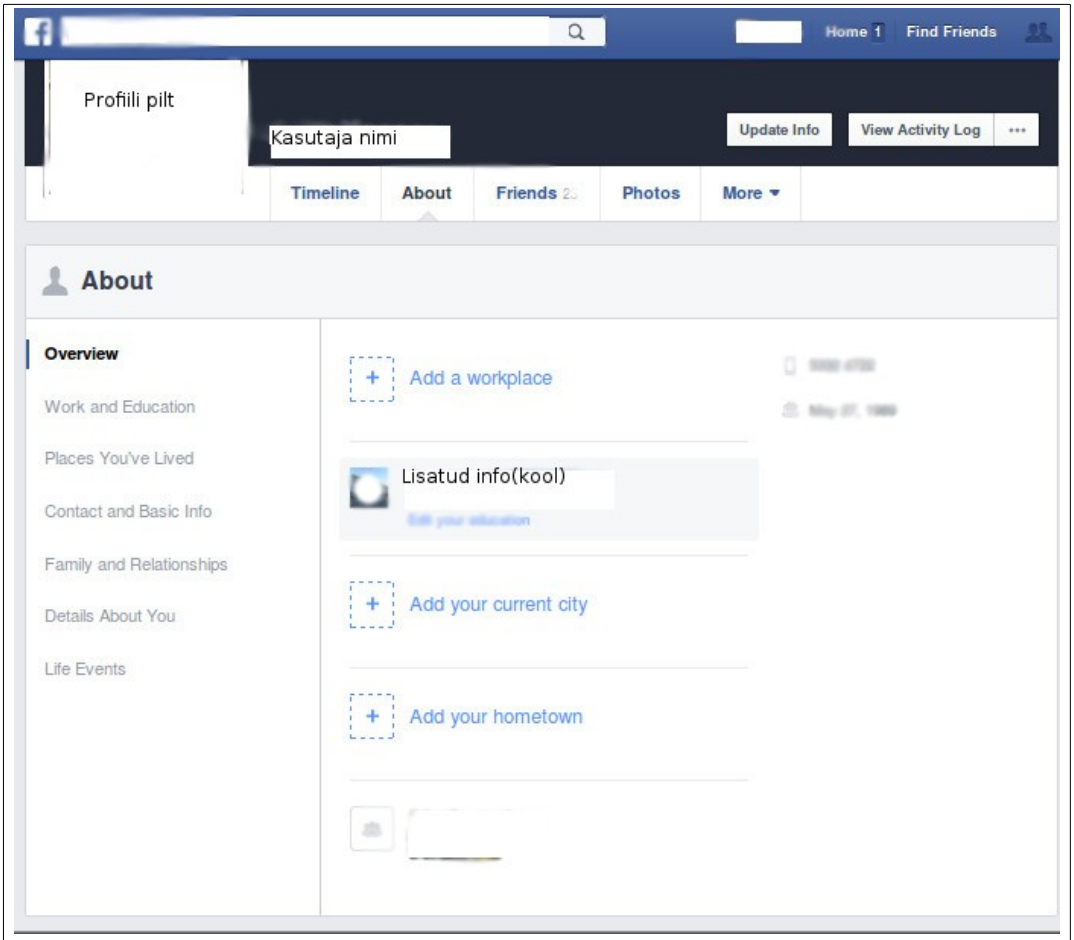
- Rakendusliidesed ehk API-d on sotsiaalvõrgu teenuste enda poolt pakutavad info jagamise teenused. Kuigi nad ei luba ligipääsu infole, mida teenusepakkuja ei ole nõus avalikult jagama, on ligipääs kiire ja lubatud andmemahud küllalt suured.

Facebooki API versioonist 2.5 alates ei ole enam võimalik automaatset infokogumise funktsiooni kasutada ilma kasutaja enda antud loata ja seega ei ole sealt võimalik infot koguda. Seega ei olnud võimalik algplaani kohane võrdlus API ja kraapimise vahel: ainuke meetod, mis on tööjõuline on kraapimine.

- Ekraanilt kraapimine toimib, tõmmates alla lehekülje lähtekoodi ja sealse info analüüsi meetodite kaudu talletamisega. Seega on võimalik saada ligipääsu vahest mitte avalikustamiseks mõeldud infole, kuid see on rakendusliidestega toimimisest aeglasem ja võib tekitada probleeme kus lehe omanik keelab kraapimisega tegelevatele IP-dele ligipääsu.

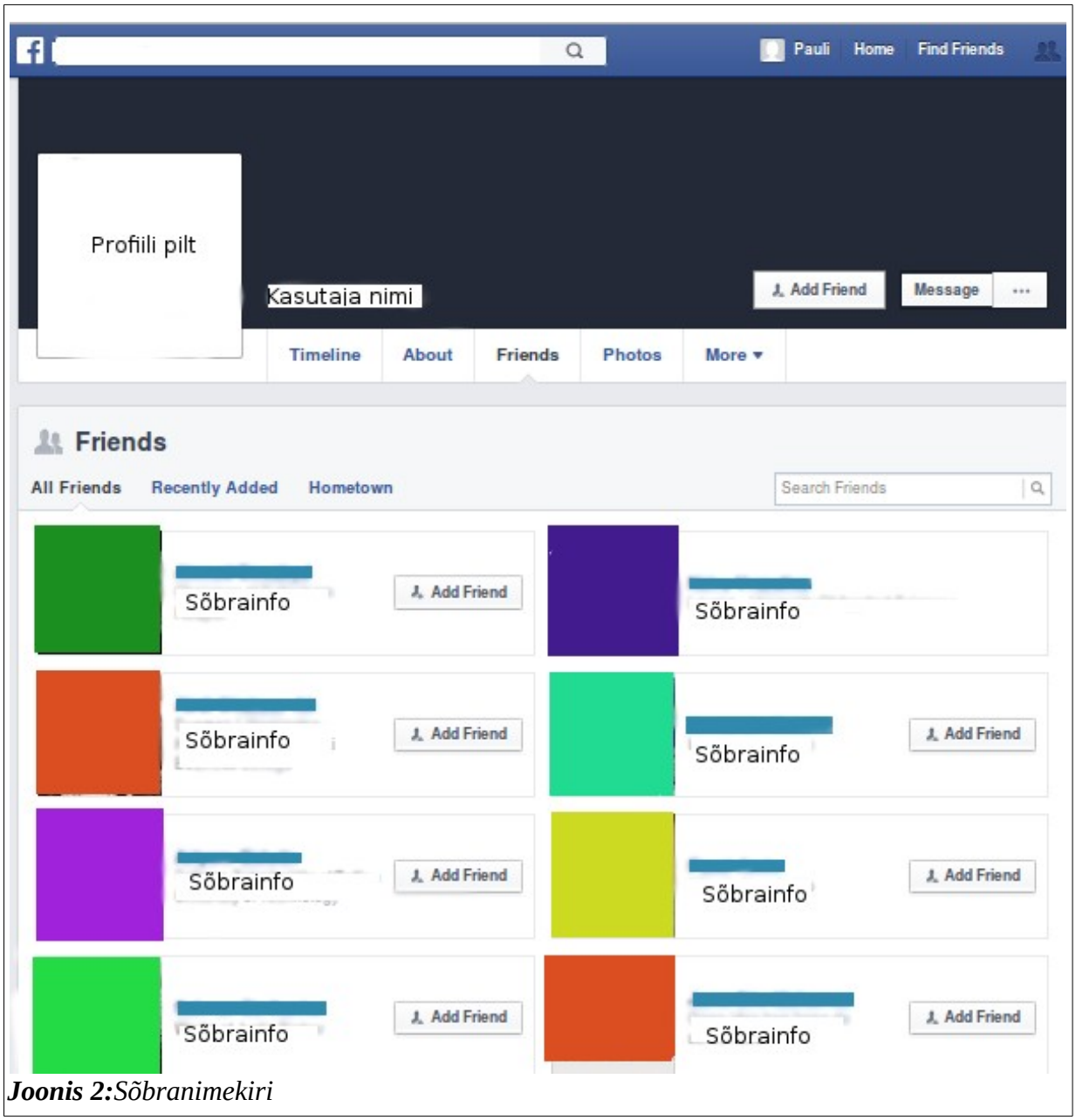
Viimasest saab mööda IP maskeerimisega kuigi seda loetakse üldjuhul lehekülje väärkasutamiseks (tingimused tihti robot.txt failis). Facebook keelas selle tegevuse oma lehel aastal 2011, kuid see ei takista seda tegelikult läbi viimast (Webster, 2015).

Esimeseks küsimuseks eduka ekraanilt kraapimise läbiviimiseks on kirjeldada, kus ja mis andmeid üles pannakse. Antud töö jaoks on soovitav neid koguda niipalju kui võimalik, nii ühe kasutaja kohta võimalikult palju kui ka võimalikult paljude eri kasutajate kohta. Üldjuhul programm, mis isikuandmete analüüsi ja kraapimisega tegeleb, on huvitatud ainult infost mis on vajalik konkreetse analüüsi tegemiseks.



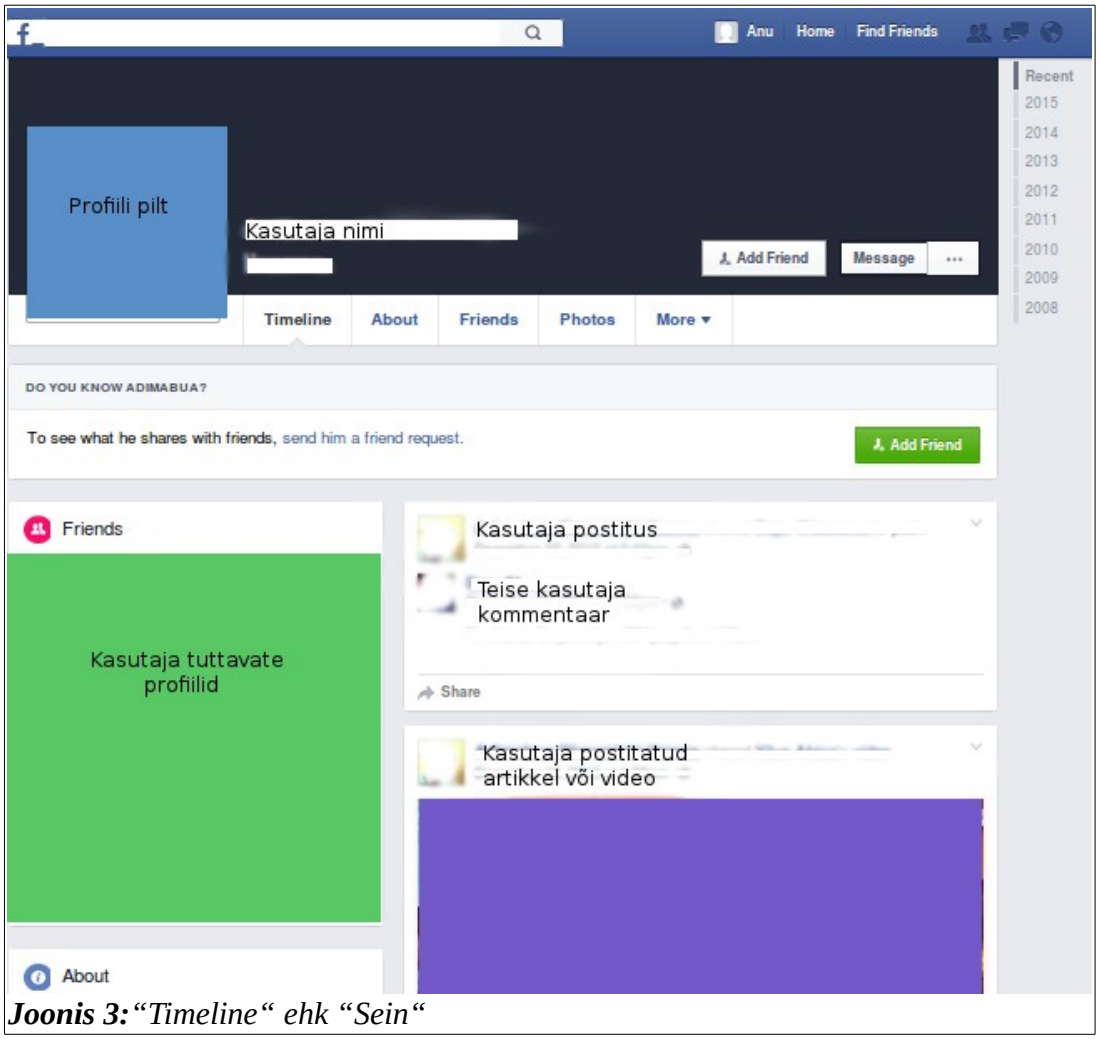
***Joonis 1:** "About" lehekülje kasutajaliides*

Esimene tüüp andmeid, mida Facebooki tavaliselt lisatakse, on isikuinfo, nt. elukoht, koolitus, sugu jne. See info on kättesaadaval „about" lehel, mis on jaotatud mitmeks osaks, tehes selle ekraanilt kraapija jaoks aeglasemaks, kuna eri lehti on vaja eraldi laadida. Samas on see konkreetne lehekülje osa parsimiseks keeruline, kuna ei ole kindel, mis andmeid on kasutaja otsustanud lisada ning nendel ei tehta lehekülje kujunduses suurt vahet. Kõige keerulisemaks teeb neile ligipääsemine see, et vaikimisi on need andmed seatud salastatuks kõikidele lehe külastajatele, kes pole kasutaja sõbrad, andes üldjuhul ligipääsu vaid kõige üldisematele andmetel („basic info") lehel. Seega võib järeldada, et Facebooki kontekstis on isiku tööajaloo, elukoha, hariduse jne andmetele kõige keerulisem kraapimise meetodiga ligi pääseda. Lehekülje allatõmbamine ja parsimine on muidugi lahendatavad probleemid ja võib-olla isegi sõbrasuhte piirangust on võimalik ümber saada, kuid selle kõige juures ei saa kindel olla kas kasutaja on üldse suvatsenud oma profiilile andmeid lisada, kuna see pole lehekülje tavakasutamiseks eriti tarvilik.



Joonis 2:*Sõbranimekiri*

Teisele lehele on kuvatud nii kasutaja sõbrasuhted kui ka ta lemmikuks valitud filmid, raamatud, arvutimängud, veebirakendused jne. Viimased on siiski vähe huvipakkuvad, kuna ei anna tingimata kasutaja endaga seostuvat informatsiooni. On kaks eri liiki profiili- suletud ja avatud profiil. Suletud profiiliga isiku sõbra nimekirja ei ole võimalik võõral näha, seega on vajalik like'de ja pildi tag-ide kaudu suhete tuletamine. Siiski on sõbrasuhted enamasti avalikud ja nende kaudu annab tuletada palju kasulikku infot. Näiteks võib avalikustatud profiiliga sõber anda vihjeid selle kohta, mis koolis või kus töötab kasutaja ja kas sõbrasuhe eksisteerib ka internetist väljaspool. Kasutaja on selle järgi tuvastatav väljaspool konkreetset sotsiaalvõrku. Sõbrasuhteid on võimalik ka kuvada Facebooki otsimootoriga („graph search") kus neid kuvatakse ilma lemmikuteta ja lisatakse info selle kohta millal sõbrasuhe loodi. Nii sõbra leht kui ka otsingu tulemused on kergesti parsitava kujundusega mis kasutajanimed selgelt välja toob.



Joonis 3:*"Timeline" ehk "Sein"*

Viimaseks kasutajate poolt loodavaks sisutüübiks on postitused, need on tekstilõigud ja pildid, mida leheküljel „sein" talletatakse ja kasutajatel omavahel jagatakse ning hinnatakse. Seina leht ise sisaldab potentsiaalselt mitme aasta postitusi ja on seega korraga kuvamiseks liiga suur, kuid postitusi saab jagada kuude kaupa kuvatavateks lehtedeks, mis on suhteliselt lihtsalt parsitavad. Postitustele lisatakse kommentaare, mille kuvamine on mõnevõrra keerulisem, kuna tarvis on kas postitust individuaalsel lehel kuvada või vajutada lingile, mis neid postitusi seina lehele laeb. Sõltuvalt kommentaaride hulgast võib kumbki lehe kraapimise keeruliseks teha. Samas on postitusi ja kommentaare analüüsides võimalik koguda huvitavat kasutajainfot ja kui kasutaja on otsustanud avalikustada postitused, kuid mitte sõbrad, on võimalik nende kaudu siiski tuletada kasutajate suhete graafi.

Töö käigus proovin skriptiga libakonto alt neid andmeid kuvada ja salvestada. Väärtuslikuks ei ole sealjuures mitte see, kas andmed on täiuslikud, vaid info selle kohta, kas midagi on lehel võõra kasutajana kättesaadav ja kui paljude kasutajate kohta korraga on võimalik sarnast infot kuvada. Määramaks, missugustele andmetele on lihtne ja millistele keeruline ligi pääseda, tuleb koostada statistikat selle kohta, mis infot kuvatakse isiku profiilil võõrastele, mis sellest ilmneb ekraanikraapides ja kuidas see erineb üle valimi.

2.1 Töövahendid

Töö läbiviimise eeskujuks oli automatiseeritud andmekraapimise skript fbstalker. Kuigi paljusid selles skriptis leitavaid võtteid on töös kasutatud, on skript tänaseks vananenud ja ei toimi Facebook'i praeguse kasutajaliidese kaudu. Selle originaalse funktsionaalsuse taastamine nõudis mitmeid uuendusi ja lõplik variant ei sarnanenud enam paljuski oma eelkäijaga.

Fbstalker on pythoni ekraanikraapimise skript, mis kuvab lehte seleniumi webdriver teegi kaudu ja tõmbab alla lehe sisu ning seejärel parsib lehekülje sisu, kasutades beautiful soup teeki. Tulemused salvestatakse nii rapordi teksti faili kui eraldi maltego andmeformaati. Selenium on eelkõige veebilehtede testimiseks loodud tehnoloogia. See on „Headless" ehk siis skriptiga juhitav brauseri tarkvara, mis kuvab lehekülgi ja viib läbi seal ette programmeerituid tegevusi. Selle jaoks rakendab ta chromedriverit, mis on Google poolt arendatud liides ja eraldi käivituv fail, mis loodud chrome ning chromium brauserite kontrollimiseks. Alternatiivselt võib selliseks tööks kasutada ka firefoxi, kuna chromega ilmnesid ka mõned stabiilsusprobleemid niipea, kui brauseris tekkisid mahukama sisu laadimisega tõrked. Lisaks, kuna skript peab lehekülje sisu parsima inglise keeles, tuleb chrome-i eesti- või võõrkeele süsteemi andmed kustutada.

Beautiful soup on pythoni teek mis parsib lehekülje lähtekoodi ja leiab sealt otsitavat infot. Selle tööpõhimõtte on jaotada lehestik html sedelite põhjal hierarhiaks, et selle sisu saaks kiiremini läbi otsida (Richardson, 2007). Tegu on andmekraapimise tüüpilise töövõttega.

Trustwave spiderlabsi tulevikuplaanides oli fbstalker mõeldud sotsiaalmõjutus (social engineering) sihtmärgi luuramiseks läbi Facebook profiili. Neid plaane ei õnnestunud vähemalt avalikult realiseerida, kuid soovitav tulevikukurss oli automatiseerida protsess nõnda, et inimsekkumist poleks tarvis, skript töötaks pidevalt, jälgides profiilil toimivaid muudatusi. Otsingute läbiviimine oli plaanis teha isiku nime järgi (mitte kasutajanime pidi). Samuti oli plaan luua punktisüsteem assotsiatsiooni tugevuse märgistamaks ja selle tarvis photo tagide, checkini, kommentaaride ja postituste ning photo like'ide analüüs. (Lee, K., Werrett, J. 2013)

Skript avalikustati 2013 aasta Hack-in-the-box konverentsil. Kasutades lehe „Graph Search" API-d sõpru, „like" (lemmikuid), „checkins" (sisse registreerimisi) ja kommentaare tuletas see sihtisikuga seotud inimesed, sisselogimise ajad, huvialad, külastatud kohad ja kuuluvus grupid. Facebooki App-ide kasutus viitas operatisooni süsteemile, mis oli kasulik sellest sõltuvate rünnaku meetodite määramiseks. Sõpradelt ja lähedastelt phishimise kaudu kogutud info põhjal oli võimalik rünnata ka sihtisikut (nt. Abikaasa infole ligipääsedes on võimalik ligi pääseda ka sihtisiku kodusüsteemidele). Info kogumise alustamiseks oli tarvis vaid profiili nime, selleks et Graph Search annaks sõbra nimekirja. Kommentaaride like'ide Check-inide kaudu saab määrata assotisatsiooni tugevuse ja jälgides millal toimuvad postitused võis tuletada ärkvel olemise ajad.(Mimoso,2013) (Kirk, 2013)

Seda takistavad tänaseks Graph API piirangud. Lisaks sõltuvus chromedriverist paneb piirid sellele, kui kiiresti skript toimib. Samuti on tegu ühe lõimega (thread) protsessis kus kõik toimingud tehakse sekventsiaalselt. Arvatavasti oleks paralleelsete protsesside abil võimalik töödelda kiiremini ja vähendada raiskuminevat aega.

Antud töö üks oluline ülesanne oli kohandada fbstalker tänapäeva facebooki kasutajaliidesega ja koostada statistika selle toimimise tulemuste kohta. Parandatud skriptivarianti saab sisestada nimekirja inimeste kasutajanimedega ja leida ka nende sõprade suhte graafi.

Antud töös on veel tähtsateks töövahenditeks TOR ja privoxy. Esimese tööpõhimõtteid on põhjalikumalt seletatud selle kodulehel (Tor: Overview, 2016), kuid selle peamine panus andtud töös on vältida IP-põhiseid turvameetmeid Facebooki poolt. Ka privoxy on proxi liides, mis filtreerib internetti lahkuvat kasutajainfot ja aitab sellega roomajat maskeerida.

2.2 Töökäigu kirjeldus

Algne fbstalkeri Skript töötleb ühe kasutaja korraga ja selle töövõime sõltub suuresti kasutaja konto usaldusnivoost. See on pärit aastast 2013 ja tänaseks see enam ei tööta.

Põhjuseks, miks algne variant fbstalker skriptist ei töötanud, on muudatused nii Facebooki API-s kui ka lehe kujunduses, mis lõhuvad seleniumi teegi võimet lehte lugeda ja sealseid andmeid töödelda.

Antud töö esimene samm seisnes vananenud variandi uue lehe laadiga kohendamises, kasutajate kohta isikliku info kogumises, skripti modifitseerimises võrkude analüüsiks mitme isiku talletamisega, graafi loomise funktsiooni taastamises ja info talletamises kasutaja postituste kohta ja sõna sageduse analüüsiks.

Töökäik: esialgselt loodud fbstalker'i vahevariant

Peale skripti esialgsete võimete taastamist tekkis küsimus, kui suure hulga kasutajate kohta on võimalik selle meetodiga infot koguda. Fbstalker skript töötas vaid ühe inimese kohta info kogumisega, kuid oli ilme, et selle funktsionaalsust võis laiendada. Selle saavutamiseks kasutati lihtsalt sisendifaili, millest skript laadis töötlemiseks kasutajanimesid.

Failist "input.txt" võeti kasutajanimed ja tõmmati seleniumi ja chrome-i kasutades alla andmed, koostades sealjuures andmebaasi sisendeid iga individuaalselt kasutajalt kogutavate andmete kohta. Seda tehti, kuna tavanimede leidmine oli liiga keeruline ja seal on liiga lai valik ning tõenäosus, et kaks isikut jagavad sama nime. Skript kasutas eelmise variandiga samu teeke ning parool ja kasutaja konto tuli sinna samuti käsitsi tekstiredaktoris sisestada.

Teistmoodi oli see, et sõltuvalt sätetest andmeid tõmmati kas rohkem või vähem alla , sätteid oli kaks: profile ja posts.

Profile (ehk käivituskäsk: python fbstalker.py [number] -profile) üritas leida isiku avalikustatuid privaatseid andmeid: töö ja koolitusajalugu, sugu ja elukohta;
kui inimene neid ise avalikuks ei ole teinud, siis neid ei näe.

Niimoodi kasutaja info leidmine oli keeruline, kuna infot tuli leida html elementides div ja span sisalduvast tekstist, kus infojagamine on valikuline ning sama laadiga html võib sisaldada mitut sorti teatud piirides kasutajainfot. See tähendab, et tuleb teksti analüüsida selleks, et leida sealse sisu liiki ja peab olema võimalik seda sisu ennetada. Näiteks võib eeldada, et tekst mis sisaldab linna nime, viitab elukohale ja sõnad OÜ ja AS töökohale.

Number näitas, mitu korda skript võttis oma väljundi uuesti sisendina. 0-iga otsiti ainult „input.txt" sisu, 1-ga otsitakse ka nende sõbrad ja 2-ga sõprade sõbrad.
Posts (python fbstalker.py [number] -posts) tõmbas alla inimese postitused ja tema seinale pandud postituste kommentaarid ning selle kaudu tuletas ka kasutaja sõprade nimekirja. See võttis aega, kui inimesel oli 900 sõpra.

Seejärel käivitatakse teine skript, mis toodab andmete põhjal graafi ja analüüsib postitusi keywords.txt alusel. Postitusi oli tavaliselt päris palju ja inimesi niigi mitu tuhat, seega oli programm üles seatud nii, et tõmbab alla ainult esimesed sada postitust. Seda sai programmis lihtsalt muuta, sealsamas, kuhu sisselogimise andmed sisestatakse.

Käivitades graph.py loodi kaks graafi myGraph ja myGraph2, esimene loodi otse sõbranimekirja analüüsides, teine loodi tuletades fotode, kommentaaride, „like“-de jne. sisust. Otsingu tulemusi sai ka näha andmebaasis aga käivitades selle skripti sai näha, kui suur oli kokkulangevus kahe otsingumeetodi vahel, samuti analüüsiti postituste sisu.

Graafidesse, mida saab näiteks kuvada programmiga nimega Gephi, kuvati arv, mis näitas, kui mitmes postituses või kommentaaris on otsitav isik kasutanud teatud võtmesõnu ehk täpsemalt, kas leidub postituse selline sümbolite jada. Lühemate sõnade puhul võib teha eksimusi, kuid seevastu on võimalik leida ka sõnu, mis näiteks lõppesid hüüumärgiga, olid erinevates käänetes jne. Meetod ei teinud vahet sellel, kas sõna ilmus mitu korda ühes postituses, see lisas +1 iga postituse kohta esimesest sajast, mis sisaldasid antud sõna.

Antud töös loodud lõplik variant

Lisaks skriptile nõuab see toimiseks teeke selenium, pygraphml, sqlite3 ja beautifulsoup4. Samuti ingliskeelsete sätetega google chrome, chrome driverit ja python ver 2.7.

Töötlemiseks on tarvis Facebooki kasutaja nime. See on nimi, mille kaudu facebook oma kasutaja andmebaasis inimesi salvestab ja on nähtav nende konto/profiiili lehe urlis. Tavaliselt on see formaadis „eesnimi.perenimi“. Selle kaudu saab tuvastada kasutaja id numbri, mille kaudu luuakse kasutaja suhetest graaf.

Uuritavate kasutajate nimed tuleb lisada lihtteksti dokumenti input.txt, kus iga rida kuulub eri kasutajale. Selline teksti fail on juba skriptiga kaasa pakitud, kuid on kasulik ainult näitena. Kasutajad mis „input.txt“-is lisatakse graafi, seega võiks neil olla vähemalt üks ühine sõprussuhe või lootus niisugust suhet leida.

Töökäigu jooksul luuakse esmalt andmete kogumiseks andmebaas.
Andmebaasi tabelid on:

- graafiandmed
  - kõrge ja madal id on omavahel seotud siis kui nende vahel on sõbrasuhe
  - järjestatus tähendab, et ühte ja sama serva ei loeta mitu korda

- metaandmed
  - kasutajanimi
  - id
  - alla tõmmatud andmete andmemaht
  - töötlemise kiirus
  - sõprade arv
  - postituste arv

- ja lõpuks tabel mis sisaldab kõigi otsitud nimede loetelu

Skripti käivitamine selles kaustas käsuga: python fbinit.py.

Selle järgi toimuvad järgnevad tegevused:

1. skript laadib sisse otsitavate kasutajate nimed failist “input.txt”
2. loob seose andmebaasiga
3. avab seleniumi ja chromedriver-iga brauseri akna ja logib sisse Facebooki keskkonda
4. hakkab nime töötlema:
    1. töötlusamm üks on tõmmata alla kasutaja sein, kus paikneb nähtamatus javaskriptis kasutaja ID
    2. kui see õnnestub ja skripti ei blokeerita, tõmmatakse alla sõprade leht, vastasel juhul skript katkestab oma tegevuse, vältimaks suuremat ajakaotust
    3. lõpuks lisatakse andmebaasi kasutaja nimi ja id koos allatõmbamisajaga ning andmemahuga, mida alla tõmmati
5. seda protseduuri korratakse kuus korda ennem, kui skript sulgeb akna ja avab uue, kuhu logitakse sisse teise kasutajakontoga

Peale selle protseduuri lõppu, kas siis, kui kõik kasutajad on alla tõmmatud või üks libakontodest on blokeeritud, võib käivitada teise skripti fbposts1.py, mis sisuliselt kordab sama protsessi, ainult tõmmates sõprade asemel alla postitusi. Skripte võib käivitada ka paralleelselt, eeldusel, et nad kasutavad eri kasutajakontosid ja kirjutavad eri andmebaasidesse.

Kui allatõmbamise tegevus on valmis, võib käivitada kaks analüüsi tarbeks loodud skripti fbparse.py ja fbposts2.py. Esimene loeb üle sõbrad allatõmmatud andmetes ja lisab selle info andmebaasi, koostades sealjuures uue „input.txt“ faili, mis sisaldab nende sõprade kasutajanimesid ja mida saab kasutada rekursiivseks otsinguks. Koostatakse kaks faili ainput ja binput, kus binput sisaldab neid nimesid, mis olid sõbrad vähemalt kahe eelnevas sisendis olnud isikuga. fbposts2.py toimib sarnaselt, lugedes üle postitused ja lisades nende kohta andmed andmebaasi. Lisaks teeb see veel võrdlust võtmesõnadega ja lisab need andmed tekstifaili „keywordscount.txt“, kus iga võtmesõna kõrval kuvatakse number selle kohta kui tihti ta seina tekstis ilmus.

Kõigi autori poolt loodud ja siinkirjeldatud programmide lähtekood on kättesaadav githubis (Lisa 1).

2.3 Arenduse käigus tekkinud probleemide analüüs

Esimeseks väljakutseks olid kujunduse muudatused. Leht toimis peale ümberkujundusi üldjoontes sarnaselt, kuid andmed olid kuvatud teiste html siltide all. See nõudis kogu parseri taaskirjutamist. Ilmselt on sellise skripti eluiga ümberkujunduste vaheline periood. Tõsisem oli lehekülgede kadumine, mis nõudis uute lehekülgede jaoks uut allalaadimise skripti. Näiteks isikuinfoleht töö- ja elukoha kohta jagati ümberkujunduse jooksul kaheks eri leheks.

Töökäigu jooksul toimus Facebooki rakendusliidesel kolm versioonivahetust: versioon 2.3 märtsis, 2.4 juulis ja 2.5 oktoobris 2015. Samuti lakkas 30 aprillil 1.0 versiooni toetus, mille tõttu kasutaja id ei ole enam rakendusliidese kaudu päritav, koodivariant, mis kolm kuud järjest oli toiminud, lakkas seletamatult töötamast ja tuli leida viis, kuidas skript saaks töötada ilma rakendusliideselt tuleva infota või seda infot leida kuskilt mujalt. Info, mida skript rakendusliideselt nõudis, oli juba väga piiratud, kuna kehtisid juba piirangud tundmatute kontode kuvamise lubadustele. Ainuke ressurss, mida lehelt otse ei tõmmatud ja mis oli töö jaoks tarvis, oli süsteemi kasutaja id, mida kasutati isikute graafi salvestamiseks. Teoreetiliselt oleks võinud seda asendada ka hashitud kasutajanimega, kuid õnnestus leida turvalünk, millega sai vana koodi kasutamist jätkata. Nimelt, kuigi kasutaja id on eemaldatud nii rakendusliidesest kui ka leheküljelt, on see siiski alles lehel paiknevas mobiilirakenduste jaoks tarvilikus koodis. Seda koodi saab ekraanikraapimise kaudu parsida ja seega on kasutaja id avalikult nähtav, kuigi ta on ametlikult lehe liidesest eemaldatud.

Tõsisemaks turvameetmeteks on Facebookil võrguroomajate vastased käitumised, mida töö läbiviimisel sai uurida. Nimelt, niipea kui skript pandi mitmekümne isiku andmeid tõmbama, saatis Facebook välja teate, et kahtlustab kontot automaatses andmekaevandamises ja võõrastele kontodele on nädal aega ligipääs keelatud . Täpsed tingimused, mille all Facebook oli võimeline robotkasutajaid tuvastama, polnud selged. Esimesel proovikontol õnnestus edukalt tõmmata alla mitmesaja kasutaja andmed ennem, kui selle tegevust piirati. Samas oli tegemist kontoga, mis tõmbas andmeid kõvasti vähem, kui seda teeks tavakasutaja, tehes ainult paarkümmend klikki tunnis piiratud paari tunni jooksul.

Võimalikud meetodid mida Facebook võis jälgimisel kasutada, et roboti käitumist märgata , on:

- Allatõmbamise maht. Kindlasti mitte ainus meetod, kuid piirangud kehtestati ainult suurel hulgal kasutajakontode allatõmbamise korral
- Allatõmbamise kiirus. Proovikonto piirati, kui see üritas alla tõmmata kuue lehekülje sisu korraga kolme sekundi intervallides.
- Sisselogimise aeg. Kuigi skript proovis ka tõmmata andmeid aeglasemalt kui tavakasutaja, piirati seda siiski. Võimalik, et jätkuvalt päeval ja öösel tegutsemine ja mitmetunnised sisselogimise ajad tekitavad robotihoiatuse Facebooki süsteemis.
- Kasutaja IP. Algsed skriptid ei maskeerinud oma kasutaja IP-d. Võimalik, et see oli süsteemis kahtlustuse all. Konto, mis käitus aeglaselt, töötas samalt masinalt, kui teine samuti skriptiga töötav konto, et paralleelselt rakendades kiirust tasa teha. Võimalik, et Facebook mõõdab nõnda ühelt masinalt tulevat liiklust.
- Käitumise viisid. Algne skript saatis välja käske rangetes ajaintervallides. Võimalik, et see andis ennetussüsteemile vihje.

Neid võimalusi silmas pidades valmis skripti variant, mis suudab tänaseks tõmmata alla tuhandeid kasutajakontosid ilma, et Facebook seda tegevust piiraks.

Esmalt vaatame, mis tingimustel libakontosid piirati. Esimest katsekontot piirati, kui skript üritas sama konto alt paralleelselt kuuest aknast andmeid tõmmata. Seejärel loodud kontot piirati sama tegevuse eest ainult kahe aknaga töötades ainult 34 minutit ennem piirangu käivitumist. Seejärel proovitud lahendust paralleelselt rakendada, mitut aeglaselt tõmbavat kontot, luhtus. Kolm kontot tõmbasid alternatiivselt 10 tundi ja kumbki umbes 30 kasutajainfot alla ennem, kui nad piirati. Samas oli ligi tund aega vahet esimese ja teiste kontode piiramise vahel, mis tähendab, et on siiski võimalik, et Facebook töötab mingi limiidi printsiibi kohaselt.

Järgmine variant, mis töötas, käitus nõnda:

- Selle asemel, et kasutada kolme kontot paralleelselt, kasutab skript 14 kontot sekventsiaalselt.
- Peale kuue kasutajainfo allatõmbamist logib skript end välja ja logib sisse teise kasutajana. Nõnda on üks konto aktiivne vähem, kui kord tunnis.
- Skript käitub pseudorandom intervallide tagant. Käske saadetakse kahe kuni viie sekundiste pausidega.
- Skript rakendab proksit ja TOR võrku maskeerimaks infot oma töömasina kohta.

Need meetmed on piisavad, et Facebooki blokeerimist vältida. Kasutades neid meetodeid, oli võimalik luua teine skript, mis rakendas kaheksat libakontot sarnasel viisil ja mis töötas samal masinal 24 tundi ilma probleemideta. Juhul kui nendest ei oleks piisanud, oleks võimalik simuleerida tavakasutajat kuni eduka tulemuseni. Näiteks oleks võinud kontod panna tööle intervallidega, et teeselda nagu oleks neist osad magamas. Lehestik nagu Facebook ei saa oma normaalkäitumise juures täielikult infole ligipääsu piirata.

Siiski oleks selle probleemi lahendamise keerukamaks teinud kaks muudatust. Esimene neist puudutab viisi, kuidas Facebook lubab oma võrgustikus kontosid luua: nimelt ei nõuta kasutajalt rohkem infot kui nende nimi, sünnipäev ja meiliaadress. Robotitele ligipääsu vältimaks on ainult üks lisameetod nimega „checkpoint“ kus peale uue konto loomist küsitakse lisaks veel kasutaja telefoninumbrit. See meetod oleks teoreetiliselt võimeline panema iga roboti, mis end võrku sisse logib maksma isikliku telefoninumbri eest (teenused, mis müüvad telefoninumbreid suurtes kogustes on muidugi ka olemas). Mistõttu on üllatav, kui harva seda takistust kasutatakse. Eriti koostööpartneritega nagu gmaili konto omanikelt ei küsitud tihti lisainfot. See andis võimaluse luua niipalju libakontosid kui tarvis, kuna Yandex, mis on Venemaa suurim otsimootor ja pakub ka meiliteenust, mis koostöös Facebookiga ei nõua oma võrgus meiliaadressi loomiseks kasutaja käest isiklikku infot ja seega on nende kahe teenuse vahel võimalik genereerida lõpmatul hulgal võltskontosid.

Teine meetod, mida Facebook kasutab oma kasutajat turvamiseks, kuid mitte võrguroomajate vastu, on geograafiline analüüs. Kui Facebooki kontole teisest riigist sisse on logitud, annab Facebook hoiatuse, et võimalik on konto vargus. Sellist tehnoloogiat saaks arvatavasti TOR teenuse vastu rakendada ja on mitmeid lehekülgi, mis küsivad TORi kasutavatelt rakendustelt lisaks turvaküsimusi. Lihtne captcha sisselogimisel oleks skripti tööd täielikult takistanud.

Viimaseks tööprobleemiks olid suured töödeldavad andmehulgad. Töö jooksul tõmmati alla 2000 kasutaja 3 kuu postitused, kokku 5,3 gigabaiti infot.

Kuna „sein“, mis sisaldab kõiki postitusi, on laadimiseks liiga suur, tõmmati alla seina jupid kuudeks jaotatult, kuid siiski põhjustas osa kuvatavast sisust süsteemi tõrkeid. Otsest tehnilist lahendust probleemile ei leitud, kuna brauseris oli piltide ja mittevajalike skriptide laadimine väljalülitatud. Üks nimi, mis tõrget põhjustas, tuli valimist lihtsalt eemaldada. Tulevastes skripti variantides tuleb selliste tõrgetega arvestada ja nende ümber automaatset käitumist planeerida.

Teine suurtest mahtudest tulenev probleem on skripti aeglus. Osaliselt saab seda lahendada nii nagu eelpool mainitud paralleelselt käivitatud skriptidega, kuid kuna postitusi on seinal vahest tuhandeid ja igaüks võib sisaldada samapalju kommentaare ning nõuda terve lehekülje allalaadimist, on selle töö piires liiga mahukas ettevõtmine. Selle asemel tõmmati alla seinal paiknevad postitused, mille tekstist peaks piisama määramaks analüüsis üldist vestlusteemat, grupikuuluvust ja hinnanguanalüüsi. Kuid suuremate tõmbamise mahtude korral oleks potentsiaalselt võimalik tuletada terve kasutaja sotsiaalvõrk kommentaari analüüsist.

3. Tulemuste analüüs

Töö jooksul loodi skripti lõplik variant. Selle töösammud on:

1. tõmmata alla sõbranimekirjad, kasutaja info, sealhulgas sugu ja postitusajalugu
2. analüüsida postitusi, luua sõnapilv
3. analüüsida sõbrasuhteid, luua graaf ja genereerida uued sisendid

Skripti kasutati 2000 kasutaja andmete allalaadimiseks. Valim koosnes kahelt Facebooki lehelt esialgselt kogutud alg.nimekirjast

- esimene - poliitiliselt neutraalne - leht oli Tallina Tehnikaülikooli Facebooki hinnangute jagamise leht,
- teine - poliitiliselt angazheeritud - oli pagulastevastane leht Facebooki ühiskonna „EI Pagulastele“ postitustest.

Mõlema lehe pealt võeti 100 suvalist kasutajat ja laaditi alla nende sõprade nimekiri. Nendest sõpradest valiti välja tuhat, eelistades neid, kes olid sõbrad rohkem kui ühe eelneva grupi liikmega. Nende isikute kohta tõmmati alla järgnevad andmed: isiku kasutaja ID Facebooki võrgus, mida rakendati hiljem graafi joonistamiseks, isiku sõbra ja lemmikute lehekülg, millest tuletati, kas kasutaja konto on avalik ja kui suur on tema sõprade hulk ja viimase kolme kuu postitused kasutaja seinal.

Tulemused olid järgnevad:

- Pagulasvastaste lehelt tuletatud grupis 1000 isikust oli suletud kontosid 270 ehk natuke üle 1/4,
- sarnasel hulgal 234 ei olnud avalikke postitusi,
- kokku tõmmati 2,6 gigabaiti andmeid.

Kasutajakontod Facebookis võivad olla kas suletud või avatud: suletud kontodel ei kuvata infot sõbrasuhete kohta ja avatud kontol, vastupidi, näidatakse sõbrasuhteid.

- TTÜ lehelt allatõmmatud sotsiaalvõrgu 1000 isikust oli 243 suletud kontoga isikut
- 295 postitusteta
- kokku tõmmati 2,7gigabaiti andmeid

- 1000 pagulusvastaste lähedaste hulgas oli 443 isikut, kelle seinal oli üle 100 sõna teksti postituse kohta (sinna hulka loetakse ka artiklite sisu ja kommentaare, kuna lihtteksti postitusi on väga vähe)

- TTÜ-st tuletatud grupis oli see arv 365

Isikuinfot kasutajate kohta on natuke keerukam kuvada, kuna inimesed lisavad seda valikuliselt ja turvasätted on rangemad. Samuti sai protsessi kiirendamiseks eemaldatud funktsioon suhete tuletamiseks, mis fotodel ning videotel kommenteerivad inimesed sõbranimekirja lisab.

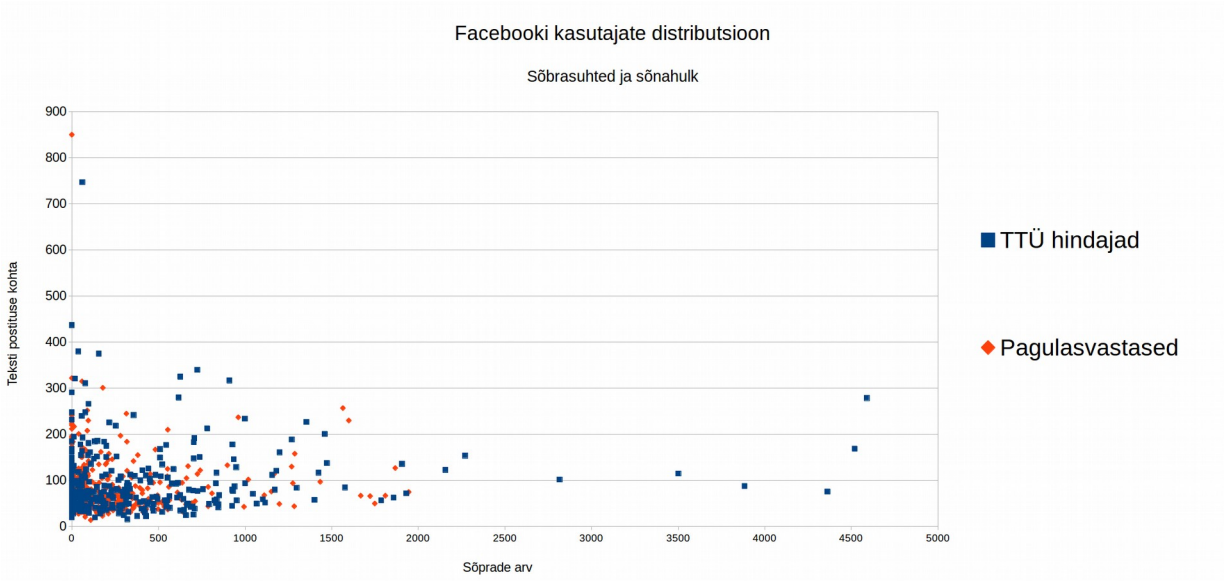
Võtmesõnade analüüsiks valisin 100 konkreetset tundlikke teemasid iseloomustavat võtmesõna, mille kasutust analüüsida: Eesti , Putin , pagulased, pagulane , pagulane, neeger, homo , kooselu , homoseksuaalsus , piir , Venemaa , venelane , venelased , Savisaar , Hitler , Stalin , seadus , assüül , Ukraina , kristlik, Jeesus , konservatiiv , konservatiivne , reform , Ilves ,

reformi, fašism , fašist , autoriteet , küüditamine , ametikoht , amet , poliitika, poliitik , poliitiline, sõda , sõja, sõdur , ateism , religioon, juut , partei , Süüria , terrorist , terrorism , quaida , politsei, riigi, riik , sõjaväe , sõjavägi , kaasmaalased , kaitse , riigikaitse , riigikaitses, Tallinn , varjupaik, valimine , valimise , Euroopa , tulumaks , käibemaks , kriis , vale , skandaal , skandaalne , pede , pankrot , riigikorraldus, pension , pensionitõus , pensionilangus , raha ,krediit , seadusemuudatus , vangistus, vang , majandus , eurotsoon , postimees , seadusandija, leibkond , sissetulek, Rootsi , usaldusväärsus, ajaloo , ajalugu , turvalisus , e-valimine, erakond , rahvas , rahvus , kultuur, keel, vägivald , kübersõda , Soome , parlamendiliikmed , minister ja NATO.

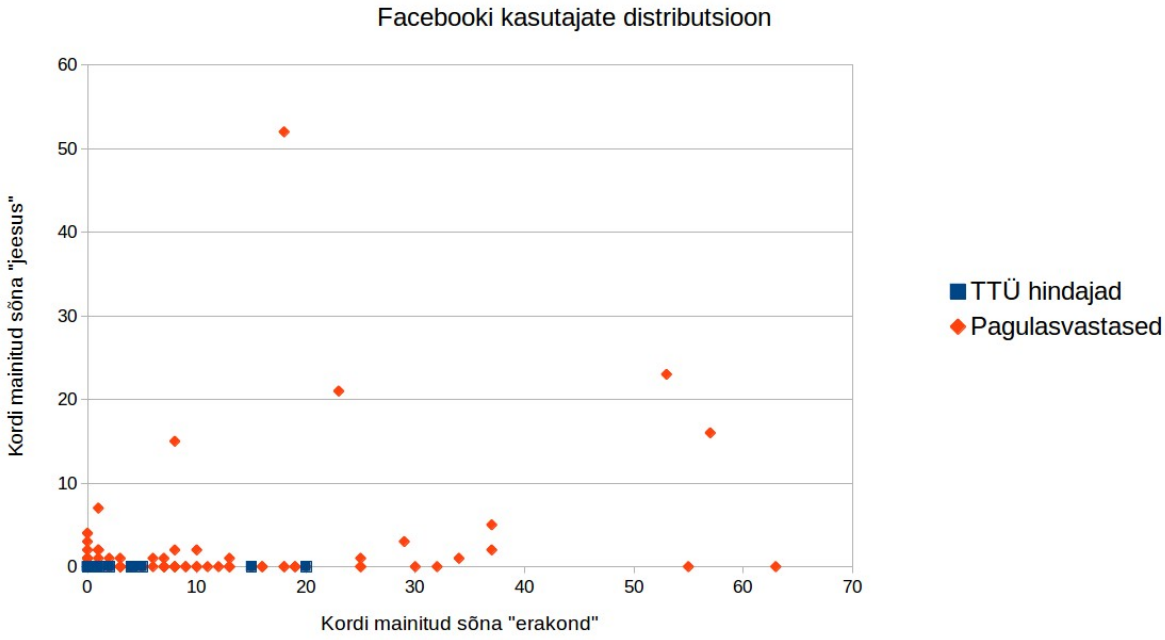
Kogutud andmeid võrreldi võtmesõnade nimekirja järgi ja kogumisel talletatud sõbrasuhetest loodi graaf (Joonis 6). Tabelis 2 on näha loend kummaski grupis kõige tihemini ilmnevatest sõnadest ja protsendimäär selle kohta, kui suur hulk sõnadest neile piirdus. Joonisel 1 ja 2 on toodud välja kahe grupi võrdlus sõnade hulga ja sõbrahulga järgi ning võtmesõnade järgi. Viimased valiti kahe grupi vahel kõige enam erinenud võtmesõnade järgi, milleks olid “erakond” ja “jeesus.”

TTÜ Facebooki lehelt tuletatud	Pagulusvastaste lehelt tuletatud
Tallinn : 2466 (0.1%) Eesti : 1138 (0.05%) minister : 656 (0.03%) NATO : 505 (0.02%) amet : 356 (0.016%) riigi : 322 (0.015%) terrorist : 256 (0.012%) vang : 254 (0.011%) keel : 235 (0.011%) Euroopa : 189(0.008%)	Eesti : 12665 (0.3%) riigi : 5368 (0.15%) Tallinn : 4640 (0.13%) Euroopa : 3379 (0.09%) raha : 3342 (0.09%) amet : 3276 (0.09%) riik : 2547 (0.07%) vale : 2494 (0.07%) seadus : 2460 (0.07%) piir : 2278 (0.06%)

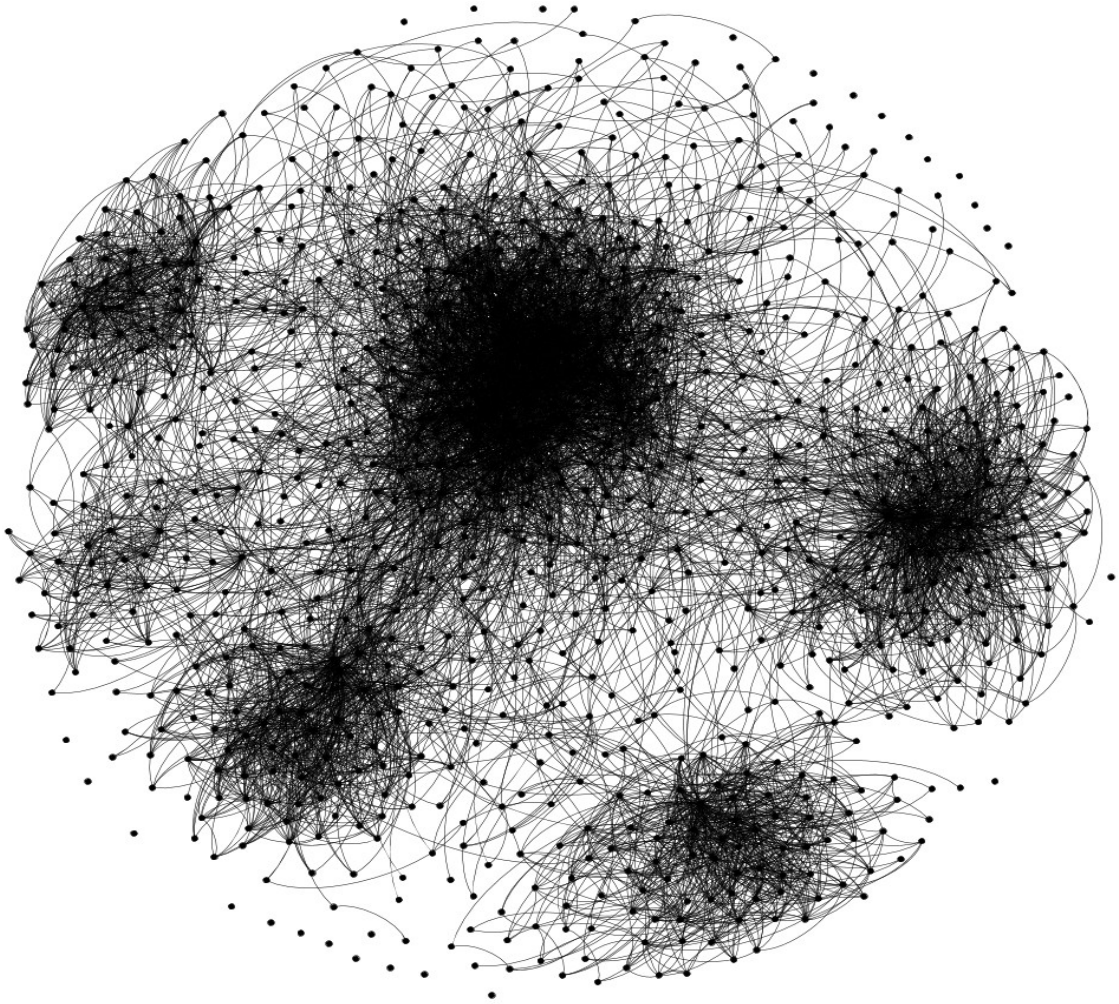
Tabel 2: *Tundlike võtmesõnade lühistatistika ülemine osa.*



Joonis 4: *Facebooki kasutajate distributsioon: sõbra- ja sõnahulk*



Joonis 5: *Facebooki kasutajate distributsioon:võtmesõnad*



Joonis 6: *Pagulasvastastest tuletatud sotsiaalvõrk: iga punkt on üks inimene, iga joon tähistab sõbrasuhet.*

4. Järeldused

Facebooki andmekraapimine libakonto kaudu on võimalik ja annab ligipääsu oluliselt rohkemale infole, kui API, mis ei avalikusta kogu ligipääsetava infot ja on seadistatud turvapiirangutega, keelamaks ligipääsu ilma kasutaja nõusolekuta. Info, mis avalikustatakse lehel sõltub profiilisätetest ja kontodevahelisest suhetest.

Kuna meetodid aeguvad lehestikel toimuvate pidevate muudatuste tõttu kiiresti, ei kata varasemad uurimistööd tänast seisu. Uuritav skript fbstalker lakkas töötamast muudatuste tõttu lehe kujunduses ja API-s.

Andmed, mis õnnestub sotsiaalvõrgustiku libakonto alt kraapimise skriptidega kätte saada, sõltuvad konto turvasätetest. Enamikelt kasutajakontodelt saab vähemalt osaliselt alla tõmmata sõbrasuhted ja teisi saab kaudselt tuletada.

Info, mida tihti avalikustatakse, on sõbrasuhted ja postitused. Ligikaudu 3/4 kontodest valimis olid seadnud Facebookis oma privaatsussätted vähemalt osaliselt avalikuks.

Sõbrasuhted on väärtuslikud, nende kaudu saab tuletada isiku identiteeti teistes sotsiaalvõrkudes. Sõltuvalt sellest kas suhe eksisteerib ka väljaspool internetti, on neid võimalik sotsiaalmõjutuse tarbeks kuritarvitada. Tõenäoline on, et sõbra kohta kehtiv info võib kehtida ka isiku kohta. Teisalt kasutatakse neid andmeid ka kuritegevuse takistamiseks ja reklaamiteenuse pakkumiseks. On võimalik teha veebiroomaja, mis inimesed Facebooki põhjal huvigruppideks jaotab ka ilma, et Facebook ise sellest tingimata teaks.

Infot saab võrrelda eraldi allikate vahel, et leida kokkulangevusi, tuvastades seega sama isiku eri sotsiaalvõrkudes. Sotsiaalmeediat võib samuti kaevandada võtmesõnade otsinguga ja leida seoseid. Analüüsi võimalused on näiteks deanonümiseerimine ja postituste suhtumisanalüüs ja grupianalüüs-

Nii need, kui ka teised sotsiaalvõrgustike andmekogumis-tegevused toimuvad nii turva- kui ka kommertspõhjustel.

Facebooki turvameetmetes on nõrkusi, mida ekspluateerides saab tavakasutaja ligi suurele andmemahule. Neid saab rakendada nt. suhtumisanalüüsi ehk „sentiment analysis“ ja võtmesõnade analüüsiks ilma, et lehe või kasutaja ise sellest tingimata teaks.

Töö raames olen loonud võrgustikke ja postitusi analüüsiva veebiroomaja, mida saab tulevikus arendada täpsemaks ja laiahaardelisemaks.

Kui soovida ennas taoliste skriptide vastu kaitsata, oleks soovitatav lokaliseerimise funktsioonid sotsiaalvõrkudes ja mobiilirakendustes välja lülitada ning privaatsussätted seada tugevaimale võimalikule tasemele.

Kokkuvõte

Eraisikute kohta info kogumine on laiahaardeline teema, mille tähtsus pidevalt kasvab. Inimesed on harjunud kasutama suhtlustehnoloogiat ja lisavad sinna vabatahtlikult aina enam võõrastele kasulikku infot, olgu siis need võõrad reklaamipakkujad või kuritegijad. Selle info kättesaamiseks ja sellest korraliku ülevaate saamiseks seisavad nii teenusepakkuja enda kehtestatud turvanõuded, kui ka inimeste enda poolt aktsepteeritavad käitumised. Nõrgad lülid turvamismeetodites annavad ligipääsu suurele hulgale infole. Kui võimalus info kogumiseks ja analüüsimiseks on olemas, ei ole keeruline koostada detailseid ülevaateid suurematest gruppidest. Info kogumine mahukates hulkades loob ka kvalitatiivseid eeliseid, näiteks on võimalik suure hulga sõprussuhetega inimesi tuvastada ilma, et nende kohta oleks isiklikule infole ligipääsu.

Töö jooksul uuriti ekraanilt kraapimise kaudu info kogumist Facebook sotsiaalvõrgustikust. Selle uurimuse tulemused on loodetavasti üldistatavad ka teistele sarnastele sotisaalvõrkudele.

Tulemuste hulgas selgus, et enamus Facebooki kasutajaid on nõus avalikustama oma profiili sisu ning sellele infole on massilisel võimalik ligipääsu saada lehekülje turvameetmetest hoolimata. Konkreetselt on Facebook keelanud oma API kaudu võõra isiku koha päringute esitamise, sealjuures ka süsteemi id päringud, kuid tänu mobiilirakenduste jaoks loodud javaskriptile on see id lehekülje lähtekoodis näha. Samuti ei ole ehtsat barjääri, loomaks lõputul hulgal kasutajakontosid veebiteenuse kaudu, kasutades näiteks Yandexit, mis nõuab konto loomiseks väga vähe isikuinfot. Neid libakontosid rakendades on võimalik leheküljele esitada suures koguses päringuid ilma, et skript blokeeritaks.

Samuti väldime IP blokeerimist TOR teenuse ja proksi kaudu töödates, mille tulemusel on kõik Facebookil paiknev avalik info piisavalt agressiivsele veebiroomajale kättesaadav. Privaatsussätetega saab kõige otsesemat ohtu vältida, kuid kuna infot saab tuletada ka sõbrasuhete kaudu ja enamus inimesi Facebookis sellist infot ei salasta, ei ole täielikku privaatsust veebilehe tavakasutuse juures võimalik garanteerida. Mõningad meetodid, kuidas võrguroomaja tööd raskemaks oleks saanud teha, oleks nõuda rangemaid tingimusi uute kontode loomisel või teha IP-de vahetamise keerulisemaks, kuna Facebooki lehe kood on võimeline tuvastama kasutaja tavapärase sisselogimise koha ja esitama lisanõudeid kinnitamaks, kui see liiga järsku muutub.

Turvameetmed on pidevas muutuses ja isegi väikesed muudatused võivad põhjustada jälgimistehnoloogias tõrkeid, mille vastu aitab ainult inimese sekkumine. Tõeline infoturve sõltub ajakohasusest ja inimtööst, kus konkreetne tarkvara aegub kiiresti. Näited selle kohta toimusid töö jooksul kaks korda kui töö käigu algul uuritavad vabavara-skriptid enam ei töötanud ja kui töö tegemise jooksul toimunud versioonimuudatus senise koodi töövõimetuks tegi. Siiski on ka vabavaralisi meetodeid info kogumiseks, nagu asukohta tuvastav cree.py tarkvara.

Sotsiaalmeedia funktsionaalsuse säilitamiseks on vaja pakkuda inimese kohta infot. Luues võltskonto sellistel lehtedel on võimalik vältida mitmeid turvameetmeid. Andmekraapimine nõuab vähest tööjõudu, kogudes samal ajal laialdast infot.

Tekstianalüüsimise teemaga haakub inimeste poolt kirjutatud kommentaaride analüüs. Statistikat selle kohta, missugust infot inimesed on valmis avaldama on kogutud sotsiaalmeedia alguspäevadest. Töö jooksul kogutud andmed viitavad muutuvatele trendidele. Info ligipääsu tundub kõige rohkem mõjutavat vaiksätestus. Inimesed ei avalikusta rohkem ega vähem infot

kui lehekülg on vaikimisi neilt nõuab. Varasemate tööde järgi võib tuletada, et enamus ei ole sätete valikust teadlik või ei hooli oma privaatsusest piisavalt ,et neid muuta.

Püstitatud probleemi võib lugeda lahendatuks, kuna arendati uudne skript ja see on töövõimeline, kuigi aeglasem, kui sarnased 2009 ja 2011 aastal ilmunud hetkel juba kõlbmatud variandid.

Edasine töö selles valdkonnas võib liikuda kahes eri suunas. Esimene suund on tekstianalüüsi meetodite parandamisele orienteeritud, eesmärgiga tuletada rohkem kasulikku infot sotsiaalvõrgu kasutusest. Teine suund on sotsiaalvõrkude enda uurimisega seotud ja arvestab lisategureid, et proovida kaardistada inimeste suhteid väljaspool internetti ja seda infot rakendada inimeste kohta info tuletamiseks. Samuti tuleb jälgida infoturbe tingimuste muutmist.

Summary

The goal of this work is finding means of data scraping from social networks to learn about information gathering software and its counter measures by testing data gathering methods.

The example project was to scrape Facebook by building technology to exploit current vulnerabilities and analysing the acquired data thus determining value of said data. This in order to answer the questions of what data is available, how it is accessed and if it can be accessed without permission and by what means? Also of interest is what sort of information such an attacker might gain via analysis of such data?

There are many past studies indicating that the Facebook network is vulnerable to data gathering. One of the open source tools used for this is the now no longer functional python script fbstalker. Restoring this lost functionality with new methods the study proved that it's still possible to scrape Facebook accounts in large numbers with a simple python script. Counter measures are circumventable and real privacy only extends to the most weakly protected friend account. Facebook network continues to be vulnerable mainly due to people's lax sense of privacy but sometimes due to ineffective security solutions.

The newly modified script was able to gain access to the data 2000 selected users. Gathering data such as posting history, friends relationship and site ID. Analysis of the data concluded that it could be used for a wide array of useful purposes such as text and group analysis as well as deducing information about people via association. Thanks to previous work in this field we also know that relationship data is a reliable means of de-anonymizing people across multiple social networks.

Kasutatud kirjandus

(Aliprandi et al., 2014) Aliprandi, C., De Luca, A. E., Di Pietro, G., Raffaelli, M., Gazze, D., La Polla, M. N., ... & Tesconi, M. (2014). CAPER: Crawling and analysing Facebook for intelligence purposes. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference* (lk. 665-669). IEEE.

(Automated Data Collection Terms.2010) Automated Data Collection Terms. (2010). [WWW] https://www.facebook.com/apps/site_scraping_tos_terms.php (05.01.2016)

(Bonneau, Anderson,. Danezis, 2009) Bonneau, J.; Anderson, J.; Danezis, G. (2009). Prying Data out of a Social Network, *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances*, lk.249,254, 20-22 , Juuli 2009.

(Compton,, Jurgens,, & Allen, 2014). Compton, R., Jurgens, D., & Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference* (lk. 393-401). IEEE.

(Dyer, 2013) Dyer, P. (2013). 50 Top Tools for Social Media Monitoring, Analytics, and Management. [WWW]http://www.pamorama.net/2013/05/12/50-top-tools-for-social-media-monitoring-social-media-analytics-social-media-management-2013/(15.04.2015)

(Webster, 2015) Webster, S. (2015). What is Scraping? [WWW] https://myhelpster.com/what-is-scraping-the-basics-for-everyone/(04.01.2016)

(Forrester, 2012) Forrester, B. (2012). Social Media Exploitation Tools: Understanding Where and How to Look. *HFM-201 Specialist Meeting on Social Media: Risks and Opportunities in Military Applications*. RTO NATO:Tallinn, Estonia. 2012.

(Forrester, 2014) Forrester, B. (2014). *Providing Focus via a Social Media Exploitation Strategy*. 18th International Command & Control Research & Technology Symposium (ICCRTS) , Alexandria, VA. U.S. 16-19 Juuni, 2014.

(Kakavas, 2011) Kakavas, Y. (2011). What is Creepy ? [WWW] http://resources.infosecinstitute.com/creepy/ (15.04.2015)

(Kirk, 2013) Kirk, J. (2013). Facebook 'stalker' tool uses Graph Search for powerful data mining. [WWW]http://www.pcworld.com/article/2056080/facebook-stalker-tool-uses-graph-search-for-powerful-data-mining.html(15.04.2015)

(Lee & Werrett, 2013) Lee, K., Werrett, J. (2013) Facebook osint: It's faster than speed dating. Hack in the Box, Kuala Lumpur, 2013. [PowerPoint slidid] http://conference.hitb.org/hitbsecconf2013kul/materials/D2T3%20-%20Keith%20Lee%20and%20Jonathan%20Werrett%20-%20Facebook%20OSINT.pdf(15.04.2015)

(Mimoso, 2013) Mimoso, M. (2013). FBstalker Automates Facebook Graph Search Data Mining[WWW]https://threatpost.com/fbstalker-automates-facebook-graph-search-data-mining/102648(15.04.2015)

(Narayanan & Shmatikov, 2009) Narayanan, A., & Shmatikov, V. (2009). De-anonymizing social networks. *Security and Privacy, 2009 30th IEEE Symposium* (lk. 173-187). IEEE.

(Reynolds et al., 2010) Reynolds, W. N., Weber, M. S., Farber, R. M., Corley, C., Cowell, A. J., & Gregory, M. (2010). Social media and social reality. In *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference* (lk. 221-226). IEEE.

(Richardson, 2007) Richardson, L. (2007). Beautiful soup documentation. [WWW]http://www.crummy.com/software/BeautifulSoup/bs4/doc/(04.01.2016)

(Salvatore et al., 2011) Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, ja Alessandro Provetti. (2011). Crawling Facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (WIMS '11). ACM, New York, USA, Article 52 , lk. 8

(Sullivan, 2014) Sullivan F. (2014). *The utilization of sock puppets in cyber intelligence operation.* Doktoritöö: NY, U.S., UTICA COLLEGE.

(Tor: Overview, 2016) Tor: Overview (2016) [WWW] https://www.torproject.org/about/overview(04.01.2016)

(Wondracek et al., 2011) Wondracek, G.; Holz, T.; Kirda, E.; Kruegel, C. (2010) A Practical Attack to De-anonymize Social Network Users, *Security and Privacy (SP), 2010 Mai IEEE Symposium lk*.223, 238, 16-19, Mai 2010.

Lisad

Lisa 1: uurimisprojekti lähtekood, saadaval:
https://github.com/i5i/fbstalk/