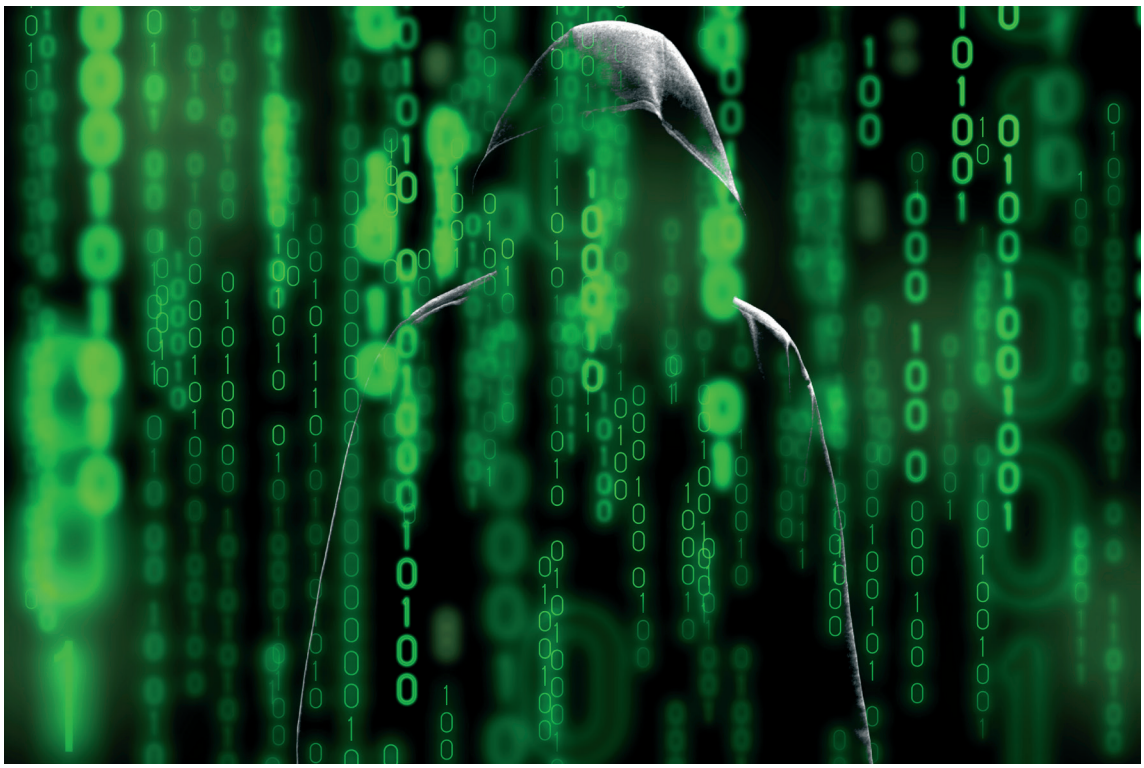
Avanonymisering i sociala medier

FREDRIK JOHANSSON, LISA KAATI, TIM ISBISTER, PETER LITSEGÅRD



FOI-R--4293--SE
ISSN 1650-1942 September 2016

Fredrik Johansson, Lisa Kaati, Tim Isbister, Peter Litsegård

Avanonymisering i sociala medier

Titel	Avanonymisering i sociala medier
Title	Deanonymization in social media
Rapportnr/Report no	FOI-R--4293--SE
Månad/Month	September
Utgivningsår/Year	2016
Antal sidor/Pages	42
ISSN	1650-1942
Kund/Customer	Försvarsmakten
Forskningsområde	1. Beslutsstödssystem och informationsfusion
FoT-område	Ledning och MSI
Projektnr/Project no	E60821
Godkänd av/Approved by	Johan Söderström
Ansvarig avdelning	Försvars- och säkerhetssystem
Bild/Cover	Shutterstock

Sammanfattning

I dagens samhälle sker alltmer kommunikation på Internet och i sociala medier. Detta gäller även vid kommunikation av mer oönskad karaktär, såsom anonyma hot på Twitter, försäljning av illegala droger på Dark Web och uppmaningar via Telegram att ansluta sig till terrorgrupper och att genomföra terrorattentat.

Vid denna typ av kommunikation används ofta användarkonton som inte explicit talar om vilken person som ligger bakom kontot, utan erbjuder någon form av anonymitet. I denna rapport presenteras tekniska lösningar som kan användas för att avanonymisera och få fram relevant information om innehavare av användarkonton på sociala medier genom metoder som författarigenkänning, aliasmatchning och användarprofilering. Tekniska möjligheter och begränsningar med att applicera de presenterade teknikerna på verklig data för brottsbekämpning och kontraterrorism tas upp. Därutöver diskuteras översiktligt eventuella etiska och juridiska problem med att använda denna typ av tekniker utanför en ren forskningsmiljö.

Nyckelord: aliasmatchning, anonymitet, användarprofilering, avanonymisering, författaranalys, författarigenkänning, sociala medier

Summary

In today's modern society, communication is to an increasing extent carried out using Internet and social media. This also holds true for communication of unwanted nature, including but not restricted to anonymous threats on Twitter, terrorism propaganda on Telegram, and selling and buying of illegal drugs on Dark Web marketplaces.

For this type of communication it is common to use social media user accounts which offer some degree of anonymity, rather than to tell explicitly who the individual behind the communication really is. This report presents various technical solutions for deanonymizing and extracting relevant information about the individuals who own social media user accounts using authorship attribution, alias matching and user profiling techniques. Technical possibilities and limitations with applying the presented techniques on real-world data for crime-fighting and counter-terrorism purposes are presented, together with a discussion on ethical and legal problems with using these kind of techniques outside the research laboratory environment.

Keywords: alias matching, anonymity, authorship analysis, authorship attribution, deanonymization, social media, user profiling

Innehållsförteckning

1 Inledning

Vid kommunikation på Internet och i sociala medier erbjuds olika former av anonymitet. I det enklare fallet tar sig detta uttryck genom fritt valbara användarnamn i sociala medieplattformar såsom Twitter, medan det på så kallade *darknets* såsom Tor aktivt understöds genom att data krypteras och skickas genom ett antal mellanliggande noder för att förhindra att utomstående får reda på vem som besöker en viss hemsida, gör ett visst inlägg eller gör en beställning på en viss marknadsplats. Möjligheten till anonymitet på Internet och i sociala medier är i grunden bra eftersom det stärker mänskliga rättigheter som åsikts- och yttrandefrihet (inte minst i icke-demokratiska samhällen) och personlig integritet, men är inte odelat positivt eftersom det också sänker ribban för näthat, anonyma hot, spridning av terrorpropaganda, utbyte och försäljning av barnpornografi och droger etc.

Även om anonyma användarkonton och tjänster används för kommunikation på webben och i sociala medier så lämnas nästintill alltid någon form av *digitala spår* som kan nyttjas för att få fram ledtrådar om identiteten bakom mer eller mindre anonyma avsändare, inte minst om dessa använder sina användarkonto frekvent över en längre tid. Ett exempel på hur avanonymisering kan åstadkommas är om anonyma användare i sina publika inlägg är oförsiktiga och lämnar ifrån sig personligt identifierbar information såsom privat mobilnummer, IP-adress, en icke-anonym email-adress eller annan information som kan knytas till individerna eller något av dessas icke-anonyma användarkonton genom samkörning av uppgifter. Det finns dock också betydligt subtilare sätt som information om anonyma användare kan läcka ut på, exempelvis speciella sätt att uttrycka sig på i text, uppgifter om vilka tidszoner de befinner sig i över en lång tid eller metadata i de bilder och dokument som användarna lägger ut.

FOI har som del av FoT-projektet ”TIA” (Teknik för Informationsfusion och Analys), det föregående FoT-projektet ”SIA” (Stödverktyg för Informationsfusion och Analys) samt Security Link-projektet ”Avanonymisering på webben” bedrivit forskning i samarbete med Uppsala Universitet för att undersöka och utveckla användning av datavetenskapliga tekniker för avanonymisering och aliasmatchning inom underrättelse- och säkerhetsinformatikområdet (främst baserat på data från webben och sociala medier, även om vissa av teknikerna även kan användas för exempelvis mobiltelefonidata).

Syftet med denna rapport är att sammanställa och redogöra för några av de främsta teknikerna inom avanonymiseringsområdet samt diskutera om och hur denna typ av tekniker kan användas av exempelvis polis- och försvarsunderrättelseorganisationer. Rapporten diskuterar också juridiska och etiska aspekter som måste beaktas innan eventuell praktisk användning samt belyser risker för den personliga integriteten om denna typ av tekniker börjar användas urskillningslöst eller för rent kommersiella syften.

1.1 Metod

Denna rapport baseras på litteraturstudier. Upplägget på dessa litteraturstudier har dock skiljt sig åt för olika delar av rapporten. Två av författarna, Fredrik Johansson och Lisa Kaati, har mångårig erfarenhet av forskning inom webbanalys, författarigenkänning och aliasmatchning. De har således genom åren hunnit bygga upp en stor kunskapsbank med relevanta vetenskapliga journalartiklar och konferensbidrag från sitt arbete inom området. För de mest intressanta och välrenommerade publikationerna har även bevakningslistor på Google Scholar lagts upp, för att erhålla notifieringar när nya artiklar inom samma ämnesområde som citerar de bevakade publikationerna publiceras. Den totala samlingen artiklar som vi på detta sätt byggt upp under åren har gåtts igenom och de, i försteförfattarens ögon, mest intressanta artiklarna för avanonymisering i sociala medier har valts ut för en noggrannare genomlysning i denna rapport.

Inom användarprofilering har författarna mindre erfarenhet, varför ett något annat upplägg använts för dessa litteraturstudier. För dessa har istället sökningar på Google Scholar och IEEEXplore efter ett antal söktermer såsom ”author profiling”, ”predicting age in social media” och ”inferring gender from posts” gjorts för att generera ett större antal träffar. Från dessa träffar har abstrakten gåtts igenom och de mest relevanta och välciterade valts ut för närmare genomgång. Referenslistorna i dessa artiklar har nyttjats för att leta fram fler relevanta artiklar. För genomgången av geolokaliseringstekniker har även ett antal befintliga litteraturstudier nyttjats för att välja ut de mest relevanta och framstående artiklarna inom området.

Författarna gör inget anspråk på att resultaten som erhållits från dessa litteraturstudier är heltäckande. Däremot är det vår uppfattning att dessa utgör ett representativt urval för att beskriva var forskningsfronten befinner sig idag kring tekniker för avanonymisering i sociala medier.

1.2 Avgränsning och läsanvisning

Denna rapport är tekniskt och metodmässigt orienterad med ett datavetenskapligt perspektiv. Policydiskussioner om huruvida exempelvis polis, säkerhetstjänst och militär underrättelsetjänst bör eller inte bör använda sig av tekniker och metoder för avanonymisering faller utanför ramarna för rapporten.

Juridisk och etisk problematik med användning av denna typ av tekniker och metoder diskuteras översiktligt, men främst problematiseras den praktiska användningen utifrån brist på träffsäkerhet och risken för felaktiga utpekanden. Rapporten går på djupet med avanonymiseringstekniker som bygger på till exempel *stylometri* och *tidsprofilering*, men behandlar utvinning av information från exempelvis bilder betydligt mer översiktligt. Olika former av *social*

ingenjörskonst, skickande av länkar som spårar besökarens IP-adress och andra typer av *aktiva åtgärder* för att försöka få individen att lämna ut personligt identifierbar information går också utanför vad som tas upp inom ramen för denna rapport. Inte heller läggs något fokus på utvinning av metadata från dokument eller bilder.

Rapporten är avsedd för allmänt tekniskt intresserade inom såväl militära som civila myndigheter inom underrättelse- och säkerhetsområdet. För att tillgodogöra sig de tekniskt djupa delarna av rapporten krävs dock förkunskaper inom framförallt allmän datavetenskap och maskininlärning. Specifika algoritmer och experiment förklaras inte alltid tillräckligt detaljerat för att möjliggöra implementation och återupprepning av experimenten.

Kapitel 2 introducerar fenomenet sociala medier och de stora mängder användargenererat innehåll som dagligen skapas samt visar på vikten av att i vissa sammanhang få fram information om individerna bakom anonyma användarprofiler. Kapitel 3 utgör kärnan av rapporten och sammanställer forskningsläget inom tre olika typer av avanonymiseringar: *författarigenkänning*, *användarprofilering* och *aliasmatchning*. I Kapitel 4 diskuteras begränsningar med dagens forskning och kunskap rörande avanonymsieringstekniker tillsammans med juridiska och etiska aspekter på praktisk användning av avanonymisering. Rapporten avslutas med Kapitel 5 där ett antal slutsatser kring användandet av tekniker för avanonymisering av användare på sociala medier presenteras.

2 Sociala medier

Sociala medier som webbforum, sociala nätverkstjänster, mikrobloggar, kommentarsfält, bloggar och vloggar (videologgar) genererar dagligen enorma mängder användarskapad data. Statistik från Domo[1] visar att på en enda minut genereras i genomsnitt över fyrahundra timmar ny film på Youtube, nästan sju miljoner filmvisningar på Snapchat, över två miljoner *likes* på Instagram och över tvåhundratusen fotodelningar på Facebook. Denna typ av statistik ger inte nödvändigtvis en helt rättvisande och heltäckande bild, inte minst eftersom de sociala medier som är populära varierar över tid, mellan länder och ålderskategorier. Många analyser av användargenererat innehåll från olika sociala medier blir utmanande på grund av volymen på den totala mängden data och den hastighet som ny data genereras med, något som populärt går under benämningen *big data*.

I dagsläget är välkända sociala nätverk i Väst t.ex. Facebook, Snapchat och Twitter, medan Kina har populära tjänster såsom Siina Weibo, Renren och QZone. I Ryssland är det istället tjänster som VK (tidigare VKontakte) och Odnoklassniki som är mest populära. På andra ställen i världen är det ofta andra sociala medier som lokalt är de mest använda.

Gemensamt för de flesta sociala medier är att det finns olika *användare* som kan skapa olika former av innehåll. Användarna identifierar sig vanligtvis med hjälp av ett användarnamn och lösenord. I vissa länder (däribland Kina) finns det krav på att användare av sociala medier registrerar sig med sina riktiga namn och sina nationella ID-kortsnummer. Även vissa tjänster som Facebook har en policy som säger att användaren måste registrera sig på tjänsten med sitt riktiga namn. I de flesta länder och med de flesta sociala medietjänster är det dock i praktiken lätt att registrera sig med påhittade användarnamn, vilket gör att användarna om de så önskar kan vara relativt anonyma. I grunden bör detta ses som positivt då det uppmuntrar yttrande- och åsiktsfrihet (inte minst i icke-demokratiska samhällen), men anonymiteten leder i många fall också till mindre önskvärt beteende. Detta kan exempelvis vara att människor använder anonymitet för att sprida näthat och att begå olika typer av brottsliga handlingar.

Återstoden av rapporten ägnas åt att sammanställa, presentera och diskutera olika former av tekniker för att avanonymisera och ta reda på mer information om innehavaren till ett anonymt konto på sociala medier. Dessa tekniker kan användas för både goda och onda syften, på motsvarande sätt som anonyma användarkonton kan användas. Etiska och juridiska aspekter tas närmare upp i Kapitel 4, huvudsakligt fokus ligger dock på teknik och den *träffsäkerhet* (eng. *accuracy*) som kan förväntas vid användning av denna typ av tekniker.

[1] https://www.domo.com/blog/2016/06/data-never-sleeps-4-0/

2.1 Data och information från sociala medier

Vilken data som finns tillgänglig vid olika typer av avanonymisering beror i mångt och mycket på vilken tjänst som är aktuell. Från vissa tjänster kan de som så önskar få ut stora mängder detaljerad data eller information, medan andra tjänster lagrar mindre information om sina användare och deras användargenererade innehåll samt är mer restriktiva med vem som kan få ut informationen. Eftersom sociala medier skiljer sig åt och förändas över tid är det tämligen meningslöst att i detalj beskriva avanonymisering för en viss specifik tjänst, annat än i syfte att exemplifiera. Istället ges i rapporten mestadels generiska exempel på data eller information som kan användas för avanonymisering.

Det är viktigt att göra en distinktion mellan vilken information som exempelvis Twitter eller ägaren till ett diskussionsforum har om sina användare och den information som folk i allmänhet kan få tillgång till. Om inte annat sägs uttryckligen utgås i denna rapport från att det är publikt tillgänglig data eller information från *öppna källor* som den som försöker utföra avanonymiseringen har tillgång till. Denna information kan i vissa specifika fall kompletteras med information från exempelvis *signalunderrättelser* eller information såsom IP-adress eller liknande som t.ex. begärs ut vid en polisutredning, men i denna rapport utgås främst från data eller information som kan fås ut från publika så kallade *applikationsprogrammeringsgränssnitt* (API:n) som tillhandahålls av den aktuella sociala medietjänsten och data som inhämtats med vanliga webbskrapningstekniker.

En grov generalisering är att de allra flesta sociala medier tillhandahåller någon form av *innehåll* och någon form av *användarinformation*. I det enklaste fallet består innehållet av någon form av text (t.ex. en post, en kommentar eller ett inlägg) och användarinformationen består av ett så kallat *alias* eller *användarnamn*. Innehåll kan på många tjänster också bestå av exempelvis bilder, video, ljud, länkar eller en kombination av olika innehållstyper, medan användarinformation utöver användarnamn exempelvis kan innehålla information om användarens e-mailadress, ålder, kön och profilbild.

Användarinformationen kan i många fall betraktas som så kallad *metadata*, men det är också vanligt att själva det innehåll som användaren genererar förses med någon form av metadata, t.ex. information om när ett inlägg har publicerats. Annan metadata på innehållsnivå kan bestå av information om vilka andra användare som har svarat på eller gillat ett publicerat innehåll. Sådan metadata kan användas för att bygga upp någon form av implicit *socialt nätverk* över vilka användare som interagerat med varandra. Många sociala nätverkstjänster tillhandahåller också mer explicit social nätverksinformation, såsom vilka användare som angivit att de är vän med en viss användare eller vilka användare som gått på samma skola.

I följande kapitel kommer exempel ges på hur data och information kan nyttjas vid olika former av avanonymisering. Vid avanonymisering av en specifik tjänst gäller det dock att undersöka exakt vilken typ av data och information som kan erhållas från denna tjänst samt att fundera på hur detta kan nyttjas vid den aktuella avanonymiseringen. De specifika resultat som kan förväntas vid avanonymisering är starkt beroende av bland annat vilken typ av data och information som kan erhållas från den aktuella plattformen.

3 Avanonymisering

Avanonymiseringstekniker kan vid en första anblick framstå som teoretiskt svårförståeliga. För att bättre förstå hur avanonymisering eller utvinning av personlig information kring en användare på sociala medier kan gå till i praktiken är det lämpligt att först begrunda några fiktiva men ändå realistiska tillämpningsfall där det kan vara aktuellt att använda sig av avanonymiseringstekniker.

- **Fall 1**: *En tonårspojke har efter en längre tids trakasserier, hat och hot på ett svenskt diskussionsforum valt att ta sitt liv. Vid en polisutredning och förhör uppkommer flera indicier och misstankar om att en viss person ligger bakom trakasserierna. Eftersom denne använt sig av anonymiseringstjänsten Tor vid inloggning på diskussionsforumet och skyddat sin hårddisk med stark kryptering så saknar polisen handfast bevisning då den misstänkte nekar. Polisen har däremot tillgång till en större mängd e-mail som personen skickat under några års tid samt delar av dennes Facebook-inlägg.*

- **Fall 2**: *Under en tid har analytiker noterat fem Twitter-konton som aktivt försöker smutskasta svenska politiker och makthavare med påhittade lögner och förtal på halvdålig engelska. Analytikerna misstänker att kontona och dess förehavanden är del av en större påverkanskampanj och vill nu försöka undersöka om det även finns kopplingar till andra Twitter-konton som inte känns till sedan tidigare samt huruvida det går att få fram någon information som tyder på att någon viss nation ligger bakom kontona.*

- **Fall 3**: *Ett företag i reklambranschen har köpt in databaser med information såsom adress, ålder och fritidsintressen om miljontals kunder från plattformarna Facebook, Pinterest och Tumblr. Företaget vill nu samköra uppgifterna från de olika databaserna för att få så bra profiler som möjligt över personerna i sina register, för att göra riktade reklamutskick baserat på personernas intressen och preferenser.*

Utifrån de tre fiktiva fallen ovan går det identifiera ett antal intressanta aspekter. För det första så går det skilja på tillämpningar där endast *en plattform* respektive *flera plattformar* eller tjänster är inblandande. Vidare går det särskilja mellan de tillämpningar där det finns personer som är kända potentiella kandidater att utgöra anonyma användare och de tillämpningar där det inte finns någon ambition att känna igen en specifik individ som ägaren till ett användarkonto. Det går också se skillnader i *antalet konton* som måste jämföras och *mängden, längden* och *kvaliteten* på information eller inlägg som finns tillgängliga. Därtill kan även olika typer av motiv till att utföra olika typer av avanonymisering skönjas. De flesta av dessa aspekter påverkar starkt vilken typ av avanonymiseringsteknik som är mest

lämplig att använda, möjligheten att säga något med tillräcklig säkerhet och huruvida det är lämpligt att använda någon form av avanonymisering.

Innan det redogörs mer i detalj för olika typer av avanonymiseringstekniker är det lämpligt att läsaren har en grundläggande förståelse för hur datorn kan användas för att automatiskt försöka säga något om huruvida två användarkonton tillhör en och samma individ, huruvida ett antal olika inlägg kan antas härröra från en och samma individ, huruvida ett konto tillhör en man eller kvinna, ung eller gammal, osv. Gemensamt för nästan samtliga tekniker som redogörs för i denna rapport är att de bygger på någon form av jämförelser mellan så kallade *särdragsvektorer* (eng. *feature vectors*). Även om detta låter avancerat är en särdragsvektor till sin natur ganska okomplicerad, där den enklast kan jämföras med en rad i Excel innehållande ett (vanligtvis relativt stort) antal numeriska värden, där varje element eller cell representerar en egenskap som exempelvis antalet förekomster av ett visst ord eller antalet ord med en viss ordlängd. Ett enskilt inlägg kan således representeras av en särdragsvektor liknande den i Figur 1 och på liknande sätt kan information om ett användarkonto representeras genom att extrahera och beräkna lämpliga särdrag.

$$< 0, \ldots, 4, 0, 1, 1, 0, 2, 0, \ldots, 0 >$$

Figur 1: Exempel på en särdragsvektor

Vid så kallad *övervakad inlärning* (vilket exempelvis många författarigenkännings- och användarprofileringstekniker bygger på) innehåller särdragsvektorerna för träningsdata också i regel ett sista element som anger vilken klass som den aktuella särdragsvektorn är ett exempel på.

Att automatiskt gå igenom inlägg och användarkonton för att skapa dessa särdragsvektorer är med tillräcklig datorkraft och en duktig programmerare i allmänhet inte särskilt svårt när all data från sociala medier som behövs är på plats, även om stora datamängder gör det mer utmanande. Det är dock i regel inte uträknandet av initiala värden i särdragsvektorn som är det komplicerade utan snarare att identifiera lämpliga särdrag att använda. När särdragsvektorerna väl finns på plats ska de också användas för klassificeringar och jämförelser. Det är främst för detta som maskininlärning och liknande metoder är relevanta. Att förklara detta ligger utanför ambitionerna för denna rapport, men förenklat kan det sägas att ju mer lika två särdragsvektorer är, desto mer sannolikt är det att inläggen eller användarprofilerna som ligger till grund för särdragsvektorerna härstammar från en och samma klass (t.ex. individ eller kön). De algoritmer som används syftar till att hitta gränserna för hur lika två särdragsvektorer behöver vara för att anses härstamma från en och samma klass, automatiskt identifiera vilka särdrag som är av störst betydelse samt att vikta olika särdrag utifrån deras relevans.

3.1 Författarigenkänning

I sin allra bredaste definition kan författarigenkänning formuleras som alla typer av försök att sluta sig till karaktäristik hos skaparen till lingvistisk data [1]. Detta är en medvetet bred definition som täcker in mycket, däribland användarprofilering som beskrivs mer utförligt i Kapitel 3.2. I praktiken avses dock med författarigenkänning vanligen ett mer begränsat problem, nämligen att:

Givet en (anonym) textmassa som är skriven av någon individ i en begränsad mängd kandidatförfattare, ta reda på vilken av kandidatförfattarna det är.

Det rör sig då inte om vilken lingvistisk data som helst utan om text. Vidare utgår problemformuleringen ifrån att det är någon av de kända kandidatförfattarna som har skrivit den anonyma texten. Kravet på att det måste vara någon medlem av en så kallad *sluten klass* av kandidatförfattare som skrivit texten håller inte alltid, utan kan ibland generaliseras till den *öppna klass*-versionen av problemet, det vill säga:

Givet en (anonym) textmassa som tros ha skrivits av någon individ i en begränsad mängd kandidatförfattare, ta reda på vilken av kandidatförfattarna, om någon, det är.

Denna version av problemet är lik den tidigare, men är i praktiken avsevärt svårare att hantera.

Författarigenkänningproblemet sträcker sig långt tillbaka i tiden, betydligt längre än vad sociala medier har existerat. Att försöka känna igen författaren till anonyma litterära verk är ett återkommande problem genom historien och Holmes redogör i [2] för många tidiga akademiska försök att med hjälp av olika statistiska metoder avgöra alltifrån författarna till olika bibliska texter till att förkasta att några omdiskuterade brev skulle ha skrivits av Mark Twain.

Ett av de tidigaste kända exemplen på särdrag för författarigenkänning föreslogs av Mendenhall [3] 1887. Mendenhall jämförde en mindre mängd författares frekvensdistributioner över ordlängd, det vill säga räknade antalet förekomster av ord med en bokstav, antalet förekomster av ord med två bokstäver, osv. Detta skulle enligt Mendenhall vara ett utmärkande särdrag som kan användas för att särskilja olika författarstilar. Flera akademiker har genom åren använt sig av statistiska test för att jämföra frekvensdistributioner för ordlängd hos anonyma texter med de hos kända författare [2].

Ett av problemen med detta är att ordlängd visat sig vara *kontextberoende*, det vill säga att en författares frekvensdistribution för ordlängd kan se ut på ett sätt när den skriver om en typ av ämne och på ett helt annat sätt när denne skriver om en annan typ av ämne. Bland annat detta har gjort att forskare genom åren föreslagit och testat en strid ström av olika särdrag för författarigenkänning, däribland distributioner över antalet stavelser per ord [4], ordlängd [5],

ordklassanvändning [6], funktionsord [7] och olika typer av ”vokabulär rikhet” [8] [9] [10].

Mycket energi har lagts ned av forskare på att ta fram och utvärdera lämpligheten av olika särdrag för författarigenkänning redan för åtskilliga år sen. Noterbart är att många av de särdrag som föreslagits genom åren också har kritiserats, då de visat sig vara för opålitliga eller ha för dålig träffsäkerhet vid användning [1].

Innebär detta att idén med att det går att få fram någon form av statistiskt robust ”skrivavtryck” för en författare som framgångsrikt kan jämföras med andra individers ”skrivavtryck” är dödfödd? Det finns mycket som talar för att det inte finns något enskilt särdrag som alltid fungerar för att särskilja individers texter. Däremot har kombinationer av många särdrag i flera studier visat sig vara väldigt kraftfullt. I många av de experiment som gjorts med kombinationer av flera särdrag, långa textmassor och ett fåtal kandidatförfattare så erhålls robusta resultat där de använda algoritmerna i de allra flesta fall pekar ut rätt kandidatförfattare. Det finns dock ett flertal kända exempel på felaktiga utpekanden som gjorts av forskare som använt sig av författarigenkänningsalgoritmer, exempelvis Qsum-algoritmen som trots detta använts i domstol [1].

Om vi lämnar den klassiska tillämpningen för författarigenkänning (där det finns endast ett litet antal kandidatförfattare och text från långa litterära verk) och istället fokuserar på tillämpningar för sociala medier så innebär detta en del nya utmaningar. Tweets, chattkonversationer etc. innebär generellt ett helt annat sätt att uttrycka sig på än vad som används i olika former av skönlitteratur. Vidare så är den totala mängden text som finns tillgänglig för en person som bara skrivit några inlägg på ett diskussionsforum avsevärt mindre att bygga robust statistik ifrån än om denne istället skrivit långa romaner. Behovet av att kunna tillämpa författarigenkänning också på kortare texter hämtade från exempelvis sociala medier är dock stort, vilket bland annat uttrycks i FBI-rapporten ”State-of-the-Art Biometric Excellence Roadmap” [11].

En välkänd algoritm för författarigenkänning som kombinerar en stor mängd särdrag och som även inkluderar särdrag avsedda att passa väl för sociala medier är Writeprint-algoritmen, utvecklad av Abbasi och Chen [12]. De har bland annat gjort experiment där de undersöker hur väl deras och andra algoritmer presterar på utmanande data från e-mail, kommentarer på en marknadsplats, ett diskussionsforum och en chatt. Resultaten från deras experiment visar på att kommentarer, inlägg på diskussionsforum samt e-mail fungerar avsevärt bättre än chattkonversationer (med en träffsäkerhet på över nittio procent för de förstnämnda och bara strax över femtio procent för chattdata vid tjugofem kandidatförfattare). De visar också att problemet blir avsevärt svårare med en ökande mängd kandidatförfattare (i medeltal runt tio procent sämre när antalet kandidatförfattare ökar till hundra).

Även om hundra kandidatförfattare är betydligt fler än vad som använts i traditionella författarigenkänningsstudier så är det långt ifrån vad som kan vara aktuellt i en del verkliga tillämpningsfall för analys av data från sociala medier. Ett antal studier med större skala har därför utförts, däribland av Koppel et al. [13] samt Narayanan et al. [14]. Dessa studier har använt sig av tiotusen respektive hundratusen kandidatförfattare, där data i båda fallen rört sig om poster från bloggar. I den senare studien har resultat erhållits där de bästa algoritmerna hittar rätt kandidatförfattare i över tjugo procent av testfallen baserat på ett fåtal bloggposter från den anonyma författaren. Denna siffra ökar till över åttio procent av testfallen när klassificeraren tillåts att avstå från att göra en gissning i de fall den inte är tillräckligt säker (vilket inträffar i ungefär 50 procent av fallen). Detta låter kanske inte så imponerande, men givet att en klassificerare som inte är bättre än slumpen i detta fall bara skulle ha rätt i ett fall på hundratusen så är dessa resultat anmärkningsvärt bra.

För den tekniskt initierade läsaren kan det vara intressant att notera att en stor mängd övervakade såväl som oövervakade algoritmer har föreslagits för författarigenkänning. Exempel på algoritmer av det förstnämnda slaget som visats prestera bra är t.ex. SVM och LDA [14], medan exempelvis Writeprints [12] och PCA [15] är exempel på mer oövervakade algoritmer som rönt framgångar. Generellt sett tycks den förstnämnda typen av algoritmer i allmänhet ge något bättre resultat på den slutna klass-versionen av problemet men har svårt att hantera öppna klass-versionen av problemet på ett bra sätt, medan den sistnämnda typen av algoritmer är lättare att anpassa till den öppna klass-versionen av problemet.

Avslutningsvis kan det konstateras att även om traditionell författarigenkänning bygger på det faktiska textinnehållet så finns det i många fall anledning att använda särdrag utvunna från innehållets metadata när det rör sig om data erhållen från sociala medier. I det generella fallet rör sig detta kanske framförallt om publiceringstid, vilken vanligtvis finns angiven för olika typer av tjänster. I specifika tillämpningar kan även annan metadata om inläggen, såsom vilken klient eller typ av webbläsare som använts för att skriva dem, finnas tillgänglig. Enbart från publiceringstiden går det att utvinna en större mängd särdrag, såsom frekvensdistribution över tid på dygnet, vardagar, helger, osv. I experiment beskrivna i [16] har vi konstruerat särdragsvektorer utifrån publiceringstiderna för de tusen mest aktiva användarna på ett irländskt diskussionsforum. Vid dessa experiment har goda resultat erhållits, där den bästa klassificeraren hittar rätt kandidatförfattare i mer än nio av tio fall trots att det rör sig om tusen kandidatförfattare och enbart tidsinformation använts för att konstruera särdragsvektorerna. Dock visar experimenten också att resultaten är starkt beroende av antalet poster, så när endast hundra poster per kandidatförfattare finns tillgängliga lyckas den bästa klassificeraren endast identifiera den rätta kandidatförfattaren i strax under två fall av tio.

Ovanstående ger förhoppningsvis läsaren en god idé om i vilka sammanhang författarigenkänning fungerar och när det inte gör det. Återgår vi till fallen som presenterades tidigare i kapitlet så är det i det första fallet som någon form av författarigenkänning kan vara tillämpligt. Det som talar emot användning är att det förutom den misstänkte personen inte är uppenbart vad som ska utgöra mängden av potentiella kandidatförfattare och att det inte heller är uppenbart att en modell upptränad på en typ av data (e-mail och Facebook-inlägg) är representativ för användning på en annan typ av data (diskussionsforumsinlägg). I detta fall skulle resultaten från en väl anpassad oövervakad algoritm med noga genomtänkta särdrag och indata kunna utgöra del av en bevisning för eller emot den misstänkte gärningsmannen. Däremot skulle det sannolikt leda till helt felaktiga slutsatser om polisen valde att rakt av läsa in de anonyma inläggen, den misstänkte gärningsmannens och några andra slumpmässigt utvalda personers inlägg i befintlig programvara för författarigenkänning. Följaktligen är det viktigt att veta precis vad som görs när författarigenkänning (likaså de andra avanonymiseringsteknikerna som presenteras i denna rapport) används.

3.2 Användarprofilering

I vissa fall finns det inte förutsättningar att på individnivå peka ut vem som har skrivit några specifika texter, exempelvis då kunskap om potentiella kandidatförfattare helt saknas. Detta gör dock inte att avanonymisering inte kan tillämpas. Liknande tekniker som använts för författarigenkänning har även tillämpats på det mer generella problem som i denna rapport benämns användarprofilering, dvs. att uttala sig om exempelvis kön, ålderskategori, nationalitet eller utbildningsnivå hos författaren till en text. Överlag kan det konstateras att detta inte är lika väl studerat som det mer precist definierade författarigenkänningsproblemet. Vi kommer i detta stycke redogöra för ett antal av de bättre studier som genomförts.

Några av forskarna som publicerat inom detta område är israelerna Shlomo Argamon och Moshe Koppel med kollegor. I ett antal artiklar har dessa studerat möjligheterna att avgöra en persons kön, ålder, modersmål och personlighet utifrån särdrag som kan utvinnas ur författarens texter. Det har också förekommit tävlingar på konferenser (däribland de välkända PAN-workshopparna[2]) där forskare försetts med träningsdata på vilken de tränat upp sina klassificerare, som sedan skickats in till tävlingsarrangörerna för att på ett oberoende sätt möjliggöra utvärdering på testdata som ingen av de deltagande forskarna haft tillgång till. En fördel med denna typ av experiment är att det blir svårare för forskarna att medvetet eller omedvetet justera algoritmerna för det specifika testdatat, vilket

[2] PAN är en årligen återkommande vetenskaplig workshop fokuserad på upptäckt av plagiarism och författarskap samt anti-socialt beteende i sociala medier.

ger en bättre indikation på hur väl användarprofilering kan förväntas fungera i praktiken.

3.2.1 Profilering av kön och ålder

Vad gäller klassificering av kön så är [17] ett typiskt exempel på tillvägagångssätt för de experiment som publicerats i forskningslitteraturen. Genom att från en datamängd bestående av böcker skrivna på engelska sortera ut texter med känd könstillhörighet hos författaren och balansera dessa så att det finns lika många texter skrivna av män som av kvinnor erhöll forskarna en slutgiltig datamängd innehållande 566 dokument. Från denna extraherades två typer av särdrag: funktionsord samt kombinationer av ordklasstaggar (exempelvis antalet förekomster av en preposition följd av ett substantiv i singular), som tillsammans utgjorde strax över tusen särdrag. Dessa lagrades i en särdragsvektor där de ingående elementen normaliserats baserat på textlängd och där även den tillhörande klassen (man eller kvinna) sparats. Efter att ha tränat upp och utvärderat en maskininlärningsalgoritm (i detta fall en variant av en så kallad Balanced Winnow-algoritm) på denna data genom användning av korsvalidering erhölls en träffsäkerhet på strax över 73 procent för funktionsorden, strax över 70 procent för ordklasstaggarna samt strax över 77 procent för kombinationen av både funktionsord och ordklasstaggar. Detta kan jämföras med slumpen som träffat rätt i 50 procent av fallen på samma datamängd.

Samma forskargrupp har i [18] gått vidare med att studera hur väl en liknande typ av metod kan användas för att klassificera både kön och ålderskategori för bloggare. Den stora skillnaden i denna nyare studie är att forskarna utöver de särdrag som diskuterats ovan även använde sig av så kallade *kontextbaserade särdrag* (exempelvis ord som har att göra med teknologi, mode och pengar). Dessa ökar algoritmernas träffsäkerhet men är med största sannolikhet inte generaliserbara nog att fungera även för andra typer av data, vilket gör dem mindre användbara för de flesta praktiska tillämpningar av intresse i denna rapport. I studien uppmättes även i detta fall ungefär 75 procent träffsäkerhet för kön när endast de *stilbaserade särdragen* (främst innefattande funktionsord och ordklasstaggar) användes, medan kombinationen med kontextbaserade särdrag gav ytterligare cirka 5 procentenheters träffsäkerhet. Samma särdrag nyttjades för klassificering av ålder, där algoritmens jobb var att skilja på tre olika kategorier: åldrarna 13-17, 23-27 samt 33-42 år. I detta experiment erhölls en total träffsäkerhet runt 76 procent när både stilbaserade och kontextbaserade särdrag användes. När enbart de stilbaserade särdragen användes blev resultaten cirka 5 procentenheter lägre. Noterbart i detta sammanhang är att algoritmen hade relativt lätt att skilja mellan den yngsta ålderskategorin och de äldre ålderskategorierna. Problemet med att skilja på de två äldre ålderskategorierna var betydligt svårare. När forskarna studerade vilka särdrag som var mest

användbara vid klassificeringen framkom det att kvinnorna oberoende av ålder i högre grad än männen använde sig av *pronomen* och *negationer* (dvs. skrev mer ”personligt”), medan männen i högre utsträckning använde sig av *artiklar* och *prepositioner* (dvs. skrev mer ”informativt”). Samma typer av särdrag som var typiskt manliga visade sig vara mest utmärkande för ålder, där användandet av artiklar och prepositioner alltså ökar med ökad ålder (men där det fortfarande i många fall är görbart att särskilja en äldre kvinna från en man).

För att få en uppfattning om hur väl klassificering av kön fungerar på svenska har FOI gjort initiala experiment på blogginlägg extraherade från ca 1500 svenska bloggar. Hälften av dessa bloggar har författats av män och andra hälften av kvinnor. Genom att använda olika kontextoberoende särdrag såsom frekvens av personliga pronomen och olika ord som har att göra med framtiden har en träffsäkerhet på ca 84 procent erhållits vid bestämning av författarens kön med hjälp av en SVM-klassificerare. Dessa initiala resultat tyder på att det finns goda möjligheter att avgöra kön även utifrån anonyma texter författade på svenska.

Fungerar de särdrag som framkommit som särskiljande mellan män och kvinnor även i mindre personlig kommunikation än just blogginlägg? Frågeställningen svaras åtminstone delvis på i en studie utförd av Cheng et al. [19]. I denna studie användes en stor uppsättning särdrag (däribland funktionsord, negationer och sådana som har att göra med vokabulär uttrycksfullhet) för att träna upp klassificerare (t.ex. SVM och AdaBoost) som försöker skilja på manliga och kvinnliga texter från olika domäner. I detta fall utgjordes domänerna av skickade e-mail (som är mer personliga) och nyhetsrapportering (som är mer formell och saklig). Studien kan kritiseras på ett antal punkter, däribland det relativt begränsade antalet författare och de obalanserade datamängderna som gör resultaten svårare att tolka. På det stora hela tyder dock de rapporterade resultaten (ca 76 procent träffsäkerhet på nyhetsrapporteringen och ca 82 procent träffsäkerhet på de skickade e-mailen) på att det finns relativt tydliga skillnader i hur män och kvinnor uttrycker sig även i korta texter som dessutom är användbara för kategorisering i mer formella sammanhang.

Utifrån ovanstående sammanställning finns belägg för att klassificering av kön och ålder baserat på anonyma texter har god potential. Dock har framförallt kön visat sig vara mer svårbestämt i vissa studier, såsom de tävlingar i författarprofilering som genomförts inom ramen för PAN[3] 2013-2016. Detta beror troligtvis på att dessa tävlingar haft än mer utmanande uppgifter, såsom att träna på en typ av data (t.ex. tweets) och utvärdera på en annan (t.ex. bloggar) och att metoderna applicerats på flera språk.

[3] http://pan.webis.de/clef16/pan16-web/author-profiling.html

3.2.2 Profilering av personlighetstyp och modersmål

Med samma typ av upplägg som de studier som redan beskrivits finns det också flera studier som försökt profilera andra typer av egenskaper hos en anonym författare, däribland modersmål och personlighet. Exempelvis har Argamon et al. [20] undersökt möjligheten att utifrån korta texter skrivna på engelska uttala sig om författarens *hemland* (i detta fall Ryssland, Tjeckien, Bulgarien, Frankrike eller Spanien) och eventuell *emotionell instabilitet*. I dessa välkontrollerade experimentuppsättningar har resultat erhållits som är väsentligt bättre än slumpmässiga gissningar, däremot finns det enligt författarna till denna rapport goda skäl att i dagsläget vara skeptisk till de flesta typer av praktisk användning av textanalys för framförallt personlighetsbedömning.

Detta betyder dock inte att det inte går att säga något om en användares personlighet utifrån vad denne delar med sig av i sociala medier. I en uppmärksammad artikel av Kosinski et al. [21] har forskarna använt sig av så kallade ”likes” på Facebook för att försöka avgöra känsliga personuppgifter såsom sexuell läggning, politiska åsikter, religion och personlighetstyp. Den upptränade klassificeraren gissade rätt på personens kön i över 90 procent av fallen och religionstillhörighet och sexuell läggning i ungefär 80 procent av fallen. Resultaten för personlighetstyp var sämre men om en tillräckligt stor mängd ”likes” kombineras med exempelvis personens texter är i det förlängningen inte osannolikt att goda prediktioner går att göra även där. I en uppföljande studie har forskargruppen till och med kommit fram till att klassificeringsalgoritmer upptränade på Facebook-likes är mer framgångsrika i att estimera en individs personlighetstyp än vad dennes närstående vänner är [22]. Ett problem med praktisk användning av denna typ av metoder, som forskarna själva inte tar upp, är att de är starkt beroende av god tillgång på uppmärkt träningsdata. Detta kan vara problematiskt då det är oklart i vilken utsträckning data som delas frekvent på sociala medier vid en tidpunkt även delas vid en annan senare tidpunkt. Dock är det två väldigt tankeväckande studier om hur mycket information om oss själva vi omedvetet kan dela med oss av, bara genom att aktivt gilla något på sociala medier.

3.2.3 Geolokalisering

Geolokalisering (även kallat *geopositionering*) innebär att försöka bestämma var en användare geografiskt hör hemma eller var denne befunnit sig vid publicering av användargenererat innehåll. Som redan nämnts så finns det möjlighet att med hjälp av särdragsvektorer och maskininlärning försöka avgöra en persons modersmål, vilket kan ses som en typ av geolokalisering. Dock finns det också flera andra typer av geolokalisering som kan bygga på allt ifrån det faktiska textinnehållet till metadata som är associerad med användarens profil till vilken plats som användarens vänner uppgett att de befinner sig på. För att bli lite mer konkreta i detta avsnitt kommer vi att fokusera på Twitter.

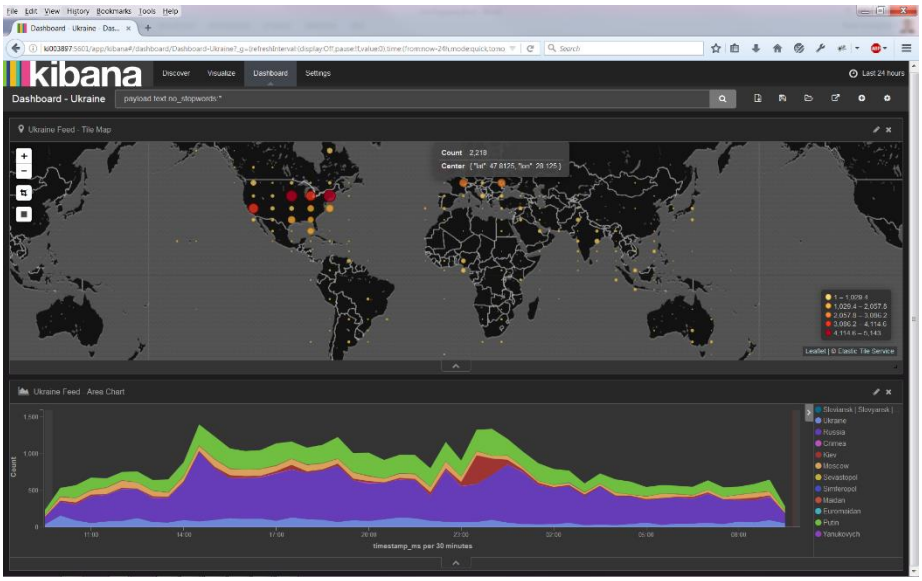
Det finns mycket som är både gemensamt och skiljer sig mellan olika sociala medieplattformar. I Twitter-fallet så finns det liksom för många andra plattformar möjlighet för användaren att aktivt geotagga sina tweets med sin aktuella position med hjälp av den GPS som finns inbyggd i de flesta moderna mobiltelefoner. I de fall detta används finns det givetvis möjligheter att över tid bygga upp en god bild av hur användaren vanligtvis rör sig, vad som är dess sannolika hemadress, jobbadress, osv.

I praktiken så är det dock relativt få användare (i litteraturen rapporteras mellan några promille upp till några få procent, beroende på studie) som regelbundet låter sina tweets förses med användarens aktuella latitud och longitud. Användare har också möjlighet att tagga sina tweets med en plats (ofta innehållande en gatuadress) som också kan extraheras tillsammans med annan metadata. I båda dessa fall krävs dock att användaren aktivt väljer att dennes tweets ska förses med tillhörande platsinformation. Detta bör alltid undersökas, men kan i de typer av tillämpningar som är av intresse i denna rapport inte förutsättas ske särskilt frekvent.

Mer vanligt förekommande är att en användare i sina tweets ger någon form av mer implicit information om var denne befunnit sig vid en viss tidpunkt, exempelvis genom att nämna ett visst landmärke, fotbollslag eller använda sig av vissa former av slanguttryck som är mer eller mindre specifika för en viss region. Tekniker för att utnyttja sådan information redogörs bland annat för i [23] [24] [25]. I några arbeten har även de URL:er som länkas till använts för att försöka avgöra i vilket land en användare befinner sig [26] (utifrån intuitionen att en svensk är mer benägen att länka till en svensk webbsida än en grek är). Dock kräver dessa typer av textbaserade tekniker i allmänhet betydligt fler tweets från en och samma användare för att ge resultat som är någorlunda tillförlitliga och användbara.

Ytterligare en typ av information som kan användas för mer naiva typer av geolokalisering är att många Twitter-användare i sin användarprofil uppger någon form av plats i fritextsformat. Denna information kan extraheras och användas för att försöka bestämma var användaren hör hemma, dock är det viktigt att komma ihåg att användaren i detta fält kan skriva vad som helst och kan uppge icke-sanningsenliga platser. I [27] uppges exempelvis att mer än var tredje användare som anger denna typ av information ger felaktiga eller påhittade uppgifter. Öppet tillgängliga kodbibliotek har dock tagits fram för att med hjälp av denna typ av information försöka geolokalisera användare, t.ex. systemet Carmen [28] som använder sig av så kallade *gazetteers* (geografiska databaser) för att försöka identifiera platser i ostrukturerad text. Forskarna bakom detta system har i sina tester visat på att de med hjälp av denna typ av metoder kan geotagga ca 20 procent av den totala mängden tweets och att de, beroende på vilken detaljeringsgrad som efterfrågas, sätter rätt geoposition i cirka 50-90 procent av fallen (där den högre siffran innebär geolokalisering på landsnivå och

den lägre motsvarar geolokalisering med en upplösningsgrad på 25km). Även om detta i regel inte är bra nog för att geopositionera enstaka användare så kan det på en aggregerad nivå ge en uppfattning om på vilka platser av intresse som det verkar ske aktuella händelser som det är värt för en analytiker att titta närmare på. Ett exempel på detta visas i Figur 2.



Figur 2: Resultat av geopositionering med Carmen av tweets som matchar specifika sökord.

Medan de geolokaliseringsmetoder som hittills redogjorts för, med undantag för innehållsanalys av själva inläggen, endast fungerar då användaren aktivt väljer att dela med sig av information om sin plats så finns det även alternativa metoder. Dessa är mer sannolika att fungera i en situation där en användare aktivt försöker hålla en mer anonym profil. En första sådan metod är att titta på vilken tid på dygnet som en användare publicerar sina inlägg på eller den tidszon som denne befinner sig i. Tweets innehåller nämligen metadata om vid vilken tidpunkt de skapats och ofta vilken tidszon användaren uppgett att denne tillhör. Även om det senare är lätt att manipulera på olika sätt kan den förväntas vara korrekt i många fall, åtminstone så länge användaren inte befinner sig på resande fot. Information om vilken tidszon en användare befinner sig i har också visat sig gå att få fram med tidsprofiler liknande de beskrivna i kapitel 3.1. En ensemblebaserad metod för geolokalisering som kombinerar ett antal enklare klassificerare innefattande estimerad tidszon, nämnda platsnamn och mest frekvent besökta plats (genererade av check-ins via Foursquare) redogörs för i [29]. Denna metod uppges av forskarna prestera bättre än tidigare föreslagna

metoder i forskningslitteraturen. När den tillämpats på tweets inhämtade från hundra användare i var och en av de hundra största amerikanska städerna så har deras framtagna algoritm uppnått en träffsäkerhet på 58 procent på stadsnivå, 66 procent på statsnivå samt 73 procent på tidszonsnivå. Dessa resultat är relativt imponerande, dock bör det has i åtanke att dessa resultat grundas på ett begränsat antal klasser (hundra städer, antalet stater som svarar mot de hundra största städerna och fyra tidszoner) och cirka tvåhundra tweets per användare. Experimenten visar att de erhållna resultaten blir sämre med ett mindre antal tweets per användare och det är ganska uppenbart att de flesta praktiska applikationer inte kan bygga på att användaren endast kan befinna sig i någon av de hundra största amerikanska städerna.

Ett starkt komplement till de innehållsbaserade och metadatabaserade metoder som redogjorts för ovan är metoder som baseras på de relationer mellan användare som finns explicit eller implicit uttryckta i olika former av sociala medier. Ett pionjärarbete inom detta område gjordes av Backstrom et al. [30] som 2010 föreslog hur en Facebook-användares vänners uppgivna adresser kunde utnyttjas för att försöka prediktera var användaren hemadress ligger rent geografiskt. Intuitionen bakom detta är enkel, du tenderar att vara vän med fler personer där du bor (eller har bott) än från andra platser. Applicerande av resultaten visar att ungefär hälften av de användare som inte uppger sin adress av integritetsskäl men har Facebook-vänner som uppgivit sina adresser ändå kan förväntas geopositioneras inom cirka 15km från sitt hem.

På Twitter existerar inte samma explicita nätverk av vänrelationer, dock har många forskare med goda resultat i kontrollerade experimentuppsättningar nyttjat mer implicita nätverk. Sådana nätverk kan byggas upp utifrån Twitter-användares följare och de konton som användaren själv följer, likheter i innehållet i de tweets de skapar samt överlapp i de platser som användarna befunnit sig på för geopositionering, se t.ex. [31]. Genom att träna upp ett så kallat dynamiskt Bayesianskt nätverk för varje användare av intresse har forskarna i [31] lyckats geopositionera några tusen användare med runt 80 procent träffsäkerhet, givet geopositioner för två eller fler av deras vänner. Dessa resultat bygger dock på att den algoritm som använts haft tillgång till relativt stora mängder historisk GPS-etablerad platsdata för användaren i fråga. Detta är sannolikt inte fallet i de flesta praktiska tillämpningsfall, vilket för denna rapport gör det mer relevant att fokusera på resultaten för en oövervakad variant av deras algoritm, där algoritmen inte haft tillgång till någon historisk platsdata för användaren vars plats vi försöker bestämma. Med tillgång till historisk och nuvarande platsdata för två av målanvändarens vänner har ca 47 procent träffsäkerhet erhållits, vilket ökat till 57 procent om detaljerad platsdata istället funnits tillgänglig för nio av målanvändarens vänner.

Ovanstående är imponerande resultat, dock kan dess praktiska konsekvenser ifrågasättas till en viss grad eftersom studien endast inkluderar vad forskarna

kallar *geo-aktiva användare*. Enligt deras definition är detta användare som skickat minst 100 GPS-taggade tweets under en månad. Detta rör sig troligtvis om mindre än en promille av samtliga Twitter-användare. De rapporterade resultaten för Flap-algoritmen som presenteras i [31], liksom resultaten för många andra nätverksbaserade geopositioneringsalgoritmer är därför svåra att generalisera till mer realistiska tillämpningar. I en kritisk jämförande studie av ett flertal sådana algoritmer som lyfts fram som state-of-the-art presenterar Jurgens et al. i [32] mer realistiska resultat som kan förväntas om algoritmerna utsätts för mer verklighetstrogna förutsättningar. Eftersom de föreslagna algoritmerna tidigare utvärderats under väldigt olika förhållanden och på olika sätt så föreslår forskarna tre gemensamma utvärderingskriterier som nio algoritmer sedan utvärderas på. Ett av resultaten är att det tydligt visas att de av algoritmerna som testats under orealistiska förhållanden (litet antal användare eller nätverk där tillgång finns till ett stort antal historiska positioner för alla användare) inte klarar av att skala upp till mer realistiska förhållanden.

Sammanfattningsvis kan det konstateras att geolokalisering såväl som andra typer av användarprofilering kan vara väldigt användbart. Vid användning för praktiska tillämpningar i verkliga världen går det dock inte att förvänta sig resultat i paritet med många av de mer välkontrollerade och begränsade experiment som rapporterats i befintlig forskningslitteratur på området. För de fall som presenterades i början på kapitlet kan olika former av användarprofilering utgöra en del av arbetssättet för att försöka utröna mer information om de personer som ligger bakom användarkontona i både det första och andra fallet, givet representativ data att träna upp klassificerare för exempelvis ålder och kön (fall 1) och modersmål (fall 2). Däremot så skulle resultatet inte kunnat användas som mer än indicier i analysarbetet.

3.3 Aliasmatchning

Problemet med aliasmatchning är starkt besläktat med både författarigenkänning och användarprofilering men bygger i allmänhet till en större grad på själva användarprofilen och tillhörande metadata än på själva innehållet. Detta gäller framförallt vid traditionell aliasmatchning (eller *user linkage* som det också kallas), som vanligtvis är inriktad på att länka samman användarprofiler där användaren inte aktivt försöker motverka sammanlänkning. I dessa enklare fall används ofta särdrag som extraherats från framförallt användarnamnet. I vissa fall används också ytterligare information såsom ålder, kön och plats. I de fall där användaren aktivt försöker motverka sammanlänkning är denna typ av information inte lika användbar, eftersom användaren då kan antas avstå från att publicera information såsom sin ålder och plats samt försöka använda sig av olika användarnamn som inte enkelt kan matchas samman. Detta gör att denna typ av aliasmatchning behöver förlita sig på mer sofistikerad analys såsom sättet som författaren uttrycker sig på i sitt användargenerade innehåll eller tidpunkten

för när inläggen skrivits. Vid sådan aliasmatchning blir därför tekniker som nyttjas vid författarigenkänning och användarprofilering högaktuella.

Oavsett om användaren aktivt försöker motverka sammanlänkning eller ej så är det av intresse att göra en distinktion mellan aliasmatchning inom en och samma sociala medietjänst (t.ex. användarprofiler på ett visst diskussionsforum) eller mellan sociala medietjänster (t.ex. Twitter och Facebook). Det första fallet kan exempelvis vara relevant då det finns intresse av att följa en användares inlägg över tid och det finns anledning att tro att denne bytt konto. Anledningar till detta kan vara att personen skapat ett extra konto för att det tidigare stängs av, fått dåligt rykte eller att ett anonymt konto skapats av anonymitetsskäl. Det andra fallet kan vara relevant då en så fullständig informationsbild som möjligt av en användare eftersträvas (eftersom en användare ofta delar med sig av en viss typ av information på exempelvis LinkedIn och en annan typ av information om sig själv på exempelvis Instagram). Detta kan exempelvis vara relevant vid kommersiella tillämpningar såsom riktad marknadsföring men också när en analytiker försöker få fram identiteten hos en användare som dold bakom ett taget användarnamn sprider propaganda och försöker rekrytera nya medlemmar till terrorgrupper.

I den mest basala formen av aliasmatchningsproblemet så har två användarkonton identifierats som att potentiellt tillhöra en och samma person. Genom att skapa särdragsvektorer som representerar dessa användarkonton så blir problemet närmast att försöka avgöra huruvida dessa särdragsvektorer är tillräckligt lika för att antas härstamma från en och samma användare.

Generaliseras detta problem något så har vi återigen ett användarkonto som vi vill extrahera ut särdrag ifrån och jämföra, men denna gång inte med ett annat konto utan med en lista eller mängd med andra användarkonton. Består denna mängd av endast ett fåtal konton så kan mer tunga och komplexa jämförelser göras, medan det om det rör sig om åtskilliga tusen eller rentav miljontals användarkonton blir viktigare med vilken beräkningsmässig komplexitet som jämförelser görs.

I det mest generella fallet av aliasmatchning finns det inget initialt konto att utgå ifrån utan problemet blir istället att utifrån en (potentiellt stor) mängd konton länka samma samtliga konton som tillhör en och samma individ. I detta fall kan det vid förekomst av miljontals konton kräva enorma mängder parvisa jämförelser, vilket för praktisk användning kräver en begränsning i vilka särdrag som används eller att endast de användarkonton som är tillräckligt lika givet enkla heuristiker blir kandidater för jämförelse av mer beräkningsintensiva särdrag.

Forskningen inom aliasmatchning är ganska spretig eftersom problemet med att matcha eller länka samman information som tillhör en och samma individ eller fenomen förekommer under ganska vitt skilda forskningsfält inom

datavetenskapen. Exempelvis är det i databassammanhang ett ganska väletablerat problem (som går under namnet *record linkage* eller *entity matching*) att försöka avgöra huruvida förekomster i olika databaser refererar till en och samma entitet eller ej. Inom databasvärlden uppstår detta problem eftersom att en individ i en databas kan beskrivas med en viss uppsättning attribut och med en viss grad av fullständighet, medan samma individ i en annan databas kan beskrivas på ett delvis annat sätt, med delvis andra typer av attribut eller andra värden. Vidare kan data vara felaktig, två olika entiteter kan ha samma eller likartade namn, osv. Liknande problematik kan exempelvis uppstå inom datafusion vid så kallad målföljning då en algoritm ska försöka avgöra om ett upptäckt mål är något som systemet redan känner till men som tidigare befunnit sig på en annan plats och därför ska uppdateras eller om detektionen ska behandlas som ett nytt unikt mål.

Vid matchning eller länkning av användarprofiler så är problematiken på många sätt likvärdig med den i databas- eller datafusionsområdet, även om den information som finns tillgänglig för sammanlänkningen av profiler vanligtvis är av annan karaktär. Den kanske mest uppenbara informationen som kan användas för aliasmatchning är själva *användarnamnen*. Att jämföra användarnamn kan i mångt och mycket jämställas med strängmatchning, för vilken en rad metoder och algoritmer tagits fram.

En av de enklare strängmatchningsalgoritmerna bygger på det så kallade *Levenshteinavståndet*, vilket kan beräknas som summan av det antal *raderingar*, *infogningar* och *substitueringar* av tecken som krävs för att transformera det ena användarnamnet till det andra. Detta avstånd kan också generaliseras för att ge olika typer av transformeringar olika kostnad, exempelvis kan raderingar av tecken göras billigare än substituering av tecken. Ett annat exempel på generalisering är att tillägg av extra tecken i slutet av användarnamnet görs billigare än i början. Ett antal sådana heuristiker kan fås fram genom att teoretiskt fundera på hur användarnamn som tillhör en och samma individ kan tänkas skilja sig från olika individers användarnamn. I [33] presenteras ett antal sådana heuristiker som sedan utvärderas empiriskt.

För att med god precision utnyttja användarnamn för att länka användare mellan olika sociala medietjänster så krävs dock i regel mer än bara några enkla heuristiker. Perito et al. [34] har i sin forskning utgått från en stor mängd användarnamn (från olika sociala medier) och nyttjat information tagen från Google Profiles, där användare listar sina användarnamn på olika sociala medier för att skapa träningsdata innehållande par av sammanhörande och icke-sammanhörande användarnamn. Med hjälp av denna data har de byggt en så kallad Markovkedja som väger in hur frekventa sekvenser av tecken varit i deras observerade träningsdata i syfte att försöka avgöra om ett par av användarnamn tillhör en och samma individ eller ej.

Zafarani och Liu presenterar i [35] metoden MOBIUS som tränar upp en klassificerare utifrån över fyrahundra användarnamnrelaterade särdrag (däribland

längd på användarnamn, frekvensen på de bokstäver som ingår i användarnamnen och exakt likhet mellan användarnamn) som utvinns från en stor datamängd med par av sammanhörande och icke-sammanhörande användarnamn. De erhållna resultaten visar på en träffsäkerhet strax över 90 procent i deras experiment, vilket kan jämföras med exakt namnmatchning och metoden föreslagen av Perito, som båda når en träffsäkerhet på strax under 80 procent. Samtliga dessa resultat kan låta imponerande och fullt tillräckliga för praktisk användning. Detta kommer dock med det viktiga förbehållet att matchning utifrån användarnamn inte kan förväntas fungera i situationer där användaren anstränger sig att vara anonym.

För de fall där likheter i användarnamn inte är tillräckligt så behövs alternativa metoder. Länkning baserat på e-mailadresser har fördelen att de är mer unika än användarnamn. Vid registrering av ett användarkonto på en ny social medieplattform är det vanligt att användaren får uppge sin e-mailadress. Även om denne använder sig av en anonym e-mailadress så är det inte ovanligt att samma e-mailadress återanvänds för registrering av olika plattformar. För de som har tillgång till denna typ av information kan det därför vara ett effektivt sätt att länka samman användarprofiler som tillhör en och samma individ. Några år tillbaka i tiden var det fullt möjligt att för olika plattformar relativt enkelt automatiskt extrahera ut stora mängder e-mailadresser som potentiellt kunde användas för länkning eller aliasmatchning i stor skala av en godtycklig angripare (se t.ex. [36]), men många företag har i olika utsträckning tätat till detta genom att vara mer restriktiva med hur information om användares registrerade e-mailadresser kan fås ut. Dock kvarstår det mycket annan information i många användarprofiler som kan användas för aliasmatchning i olika utsträckning.

I [37] presenteras exempelvis metoden HYDRA för att matcha samman användare mellan olika sociala medieplattformar. Denna metod kombinerar på ett intressant sätt en rad olika typer av information, innefattande *användarprofilinformation*, *ostrukturerat användargenererat innehåll* och *användarens beteende*. Vid storskaliga realistiska experiment på flera miljoner användare har resultat erhållits som visar att den kombinerade HYDRA-metoden får signifikant bättre resultat än både MOBIUS som diskuterats ovan och ett antal alternativa metoder. Detta är i sig inte förvånande då HYDRA väger in mer information, men det är noterbart att utvärderingen av metoden visar på att den i snitt korrekt länkar samman mer än 90 procent av de användarkonton som hör till en och samma användare samtidigt som den gör felaktiga sammanlänkningar i mindre än ett fall på tio. Därmed finns det redan idag metoder som fungerar bra även under realistiska förhållanden och utan tillgång till unika identifierare såsom e-mailadresser.

Dock finns det utrymme att lösa de mer utmanande delarna av aliasmatchningsproblemet på ett bättre och mer noggrant sätt än vad HYDRA och liknande metoder gör, speciellt rörande de användare som bryr sig om att ens

olika konton inte ska kopplas samman (och som troligtvis står för en icke obetydlig del av de konton som HYDRA missar). Detta gäller inte minst användning av algoritmer som bygger på stylometri och tidsprofiler i enlighet med vad som presenterats i Kapitel 3.1 och Kapitel 3.2. I forskning från FOI och Uppsala Universitet har vi genom att kombinera tidsprofiler och stylometri för aliasmatchning på data från ett diskussionsforum fått resultat som pekar på att vi vid oövervakad aliasmatchning av femhundra individer som var och en har två olika användarkonton kan nå 100 procent träffsäkerhet, vilket vid fyratusen användare reducerats till cirka 75 procent träffsäkerhet [38]. Vid användning av övervakad inlärning där aliasmatchningen definieras som ett binärt klassificeringsproblem har vi för ett tusental användare nått en träffsäkerhet på högt över 90 procent för länkning av användarkonton inom ett diskussionsforum och inom Twitter [39]. För att dessa metoder ska fungera krävs dock att det finns relativt stora mängder data att bygga tidsprofilerna respektive de stylometriska avtrycken ifrån. Användning vid riktigt storskaliga förhållanden (miljontals konton) har inte utvärderats, men lär inte fungera i närheten av så bra och skulle dessutom ta åtskilliga dagar att bygga upp klassificerar för. För storskalig användning kan avsevärt sämre resultat än motsvarande resultat för HYDRA förväntas, men styrkan med stylometri och tidsprofiler är att de bygger på särdrag som är betydligt svårare att manipulera än exempelvis användarnamn.

Avslutningsvis finns det i många sociala medier också ett antal andra typer av mer specifik information som på olika sätt kan nyttjas vid aliasmatchning. Ett sådant exempel är sociala nätverk i form av explicita vän-nätverk på Facebook eller mer implicita nätverk såsom de som kan fås fram utifrån vilka användare som nämnt varandra på Twitter. I vissa fall kan överlapp i sådana nätverk vara användbara för aliasmatchning, vilket diskuteras mer ingående i [40]. På samma sätt är information utvunnen ur bilder något som kan vara användbart för aliasmatchning. Preliminära resultat från Bertini et al. [41] tyder på att individuella hårdvarudefekter hos mobiltelefonkameror leder till systematiskt brus som kan nyttjas för att länka samman och särskilja uppladdade bilder som tagits med en och samma kamera från bilder som tagits med olika kameror, trots att olika sociala medietjänster komprimerar bilder då de laddas upp på tjänsten. På sikt kan detta bli intressant, även om det i dagsläget är tveksamt om de preliminära resultaten skalar upp med antalet kameror (i denna studie har enbart fem olika mobiltelefonkameror ingått).

I de tidigare exemplen på användningsområden så skulle det i fall 3 (företaget som köpt in databaser med information om kunder från olika sociala medieplattformar) sannolikt vara möjligt att använda användarnamn och eventuella uppgifter om e-mailadresser tillsammans eller var för sig för att med hög precision och träffsäkerhet länka samman merparten av de konton som tillhör en och samma användare.

4 Diskussion

Begränsningar med avanonymisering som förts fram i föregående kapitel är att det finns praktiska problem såsom att träffsäkerheten sällan är hundraprocentig och att vissa algoritmer inte skalar till den volym av data och konton som en del faktiska tillämpningar kräver. I detta kapitel kommer ytterligare begränsande faktorer att diskuteras, däribland tillgång till representativ data från vilken generaliserbara modeller kan byggas, algoritmernas språkberoende samt risken för att någon utnyttjar kunskap om metodernas uppbyggnad för att åstadkomma olika typer av vilseledning. Utöver detta diskuteras även praktisk användning av avanonymiseringstekniker utifrån ett juridiskt och etiskt perspektiv.

4.1 Eventuella problem med avanonymisering

Teknikerna som presenterats i Kapitel 3 är kraftfulla och har inte minst i forskningsexperiment visat sig ha potential att fungera väl för många typer av tillämpningar. Några av de svagheter som redan belysts är att det många gånger krävs ganska stora mängder användargenererat innehåll från en anonym användare för att denne ska avanonymiseras med god träffsäkerhet samt att teknikerna tenderar att fungera sämre ju fler användare som behöver jämföras. Det finns dock fler problem, där vi i detta avsnitt valt att studera några av dem närmare.

Det första sådana problemet är att en stor andel av den forskning som gjorts på avanonymisering utförts på texter skrivna på engelska. Även om det finns undantag så är studier på exempelvis svenska, ryska, kinesiska och arabiska betydligt mer sällsynta. Metoderna som sådana bygger nästan alltid på olika former av särdragsvektorer och särdragen går i de allra flesta fall att anpassa till andra språk relativt lätt, dock med förbehållet att de språkteknologibibliotek som krävs för att utvinna de önskade särdragen såsom lemmatiserade ord, par av ordklasser etc. inte är lika utvecklade som för engelska och därmed inte fungerar riktigt lika bra. När särdragsvektorerna väl finns där så kan precis samma metoder användas för att träna upp klassificerare som för engelska, men då detta i stor utsträckning inte gjorts så är det en öppen forskningsfråga huruvida det är lika svårt, lättare eller svårare att exempelvis försöka bestämma en individs ålder eller personlighet om denne skriver på franska, jämfört med om den skriver på engelska. Detta gör inte att praktisk användning av avanonymisering för exempelvis svenska förhållanden bör avstås ifrån, men däremot finns i dagsläget en något större osäkerhet rörande resultaten jämfört med engelska texter.

En annan frågeställning värd att belysa är risken eller möjligheten för att någon medvetet påverkar den information som ligger till grund för utvunna särdragsvektorer på ett sätt som inverkar på avanonymiseringsresultaten. Att utelämnade eller minskade digitala spår gör avanonymisering svårare eller rent

av omöjlig är en sak (t.ex. användare som ser till att inte fylla i någon riktig information i sina användarprofiler och aktivt minimerar mängden metadata som kan användas), men går det att medvetet lura algoritmerna för att försöka sätta dit någon annan? Om vetskap finns kring vad en avanonymiseringsmetod bygger på går det givetvis rent hypotetiskt att samla in information om en användares användarnamn, övrig användarprofil och tidsprofil för att sedan skapa nya användare med liknande karaktäristik och beteende för att ge sken av att ens användarkonto tillhör den andra personen. Vad gäller anpassning av sitt stylometriska skrivavtryck för att försöka *imitera* en annan persons skrivsätt är det mer oklart om det går.

Till stora delar är algoritmers motståndskraft mot imitation eller så kallade ”*framing attack*s” ett outforskat område, men några initiala studier har utförts av forskare från Drexel University. I [42] har fyrtiofem frivilliga personer (som inte har någon kunskap om författarigenkänning) fått i uppgift att försöka efterlikna det skrivsätt som en känd författare med ganska utpräglat skrivsätt har. Studiens upplägg kan ifrågasättas på ett antal punkter, men forskarnas slutsats är att imitationsförsök riktade mot väldigt enkla former av författarigenkänningsalgoritmer lyckas i hög utsträckning (under gynnsamma förutsättningar med endast fem kandidatförfattare i över 60 procent av fallen). Mer avancerade algoritmer som kombinerar ett större antal särdrag är mer motståndskraftiga, men under de mest gynnsamma förutsättningarna lyckas imitationen i strax över fyra av tio fall. I [43] har skolade imitatörer försökt efterlikna Ernest Hemingways och William Faulkners skrivstil med ännu bättre framgång. Detta talar för att lyckad imitation av en persons stilistiska uttryckssätt inte kan bortses ifrån vid användning av författarigenkänning. Dock presenterar forskarna också resultat som tyder på att förekomsten av denna typ av imitationsattacker faktiskt kan upptäckas med liknande angreppssätt som de tekniker som redan presenterats i denna rapport.

Av ovanstående resonemang är det viktiga budskapet att ta med sig att det är viktigt att ha en förståelse för hur algoritmer för olika typer av avanonymisering fungerar och att det sällan går att blint lita på de resultat som användning av sådana system eller algoritmer ger. En kunnig människa måste vara med i loopen och fundera över huruvida det finns risk att de data som har använts för att träna upp en modell är representativa nog att användas i det aktuella fallet (om data är av rätt typ, om modellen är upptränad på ett tillräckligt stort urval etc.) samt om det finns skäl att tro att någon medvetet försökt styra utfallet i en specifik riktning.

4.2 Juridiska och etiska perspektiv på avanonymisering

Avanonymisering handlar inte bara om teknik utan innefattar också viktiga juridiska och etiska frågor. Är det lagligt att ladda ned information från olika typer av sociala medier i syfte att avanonymisera en eller flera användare med hjälp av automatisk databearbetning och är det etiskt försvarbart? Som så många andra frågor rörande etik och juridik har detta inget enkelt och entydigt svar, utan för att uttrycka sig lite vagt så beror det antagligen på vilka omständigheterna är. Följande beskrivning bygger till stora delar på kunskap som inhämtats genom samtal och workshops med såväl IT-rättsjurister som forskare inom etikområdet.

En första reflektion som kan göras är att avanonymisering som bygger på data och information som hämtats in från sociala medier är ett exempel på automatiserad bearbetning av något som i de allra flesta fall innehåller någon form av personuppgifter. Följaktligen så gäller för svenska företag och myndigheter att den så kallade personuppgiftslagen blir aktuell, vilken är en svensk implementering av EU:s dataskyddsdirektiv 95/46/EG. Redan detta gör det komplicerat eftersom det dels i dagsläget (2016) finns vissa skillnader i hur personuppgifter får behandlas beroende på om det rör sig om försvarsunderrättelsetjänst, polisiära utredningar eller något annat, men också genom att nuvarande personuppgiftlag den 25 maj 2018 kommer att ersättas av en ny EU-förordning om dataskydd. Detta kommer på olika sätt strama upp regelverket ytterligare, exempelvis i form av att den nuvarande så kallade missbruksregeln som tillåter behandling av ostrukturerad data troligen kommer att försvinna. Med anledning av detta blir rådande rättsläge alltmer invecklat, dock kommer vi i den följande diskussionen utgå från syftet och den uppfattade intentionen med gällande och kommande lagstiftning, snarare än specifika paragrafer.

För att bearbetning av personuppgifter ska vara godtagbart i lagens mening så går det lite grovt att säga att det (1) s*ka finnas ett informerat samtycke från individen vars personuppgifter bearbetas*, alternativt så ska det (2) *finnas ett väldefinierat syfte där det ska ske en intresseavvägning mellan den enskilde individens intresse och avanonymiserarens intresse*. Vid forskning rörande avanonymisering blir informerat samtycke ofta i praktiken en omöjlighet. Ett väldefinierat forskningssyfte kan dock förväntas väga tyngre än individens intresse (givet att forskaren gör allt för att minimera inverkan på enskilda personers integritet, se även diskussionen kring etiska överväganden). För ett kommersiellt syfte kan denna avvägning i de allra flesta fall förväntas falla i den enskilde individens favör, varför kommersiella aktörer i vår tolkning av lagens mening alltid bör ha ett informerat samtycke innan någon form av avanonymisering (såsom sammanlänkning av användarprofiler för mer riktad marknadsföring) kan ske. Vad gäller syften som ligger i nationens intresse i form

av rikets säkerhet etc. så finns det särskilda skrivningar och paragrafer för detta, vilka kunniga jurister bör titta närmare på för att göra mer professionella bedömningar. Författaranalys av olika slag har dock redan tidigare använts i rättegångssammanhang [1], bland annat i Storbritannien, vilket talar för att det åtminstone i specifika fall går att använda sig av rent juridiskt.

En annan lagstiftning som kommer i fråga vid nedladdning av material från olika sociala medier är den om upphovsrätt. Upphovsrättslagen gör det problematiskt för forskare att publicera datamängder som laddats ned från sociala medier (i syfte att exempelvis träna upp klassificerare för avanonymisering), däremot bör frågan om upphovsrätt inte vara ett problem för ”eget bruk”. Ett större problem i detta sammanhang är de så kallade *användarvillkor* eller *terms of services* som användning och nedladdning av data från olika sociala medietjänster vanligtvis kräver att användaren accepterar. Dessa är i lagens mening giltiga och bör läsas igenom noga innan någon form av analys av sociala medier (inklusive avanonymisering) påbörjas baserad på data från en viss specifik plattform.

Sett från ett etiskt perspektiv bör forskningsetik separeras från användning av avanonymisering i praktiken eftersom detta är två olika saker och där den senare kan ses som betydligt mer problematisk. Vid kommersiell avanonymisering är det intressant att även om informerat samtycke i juridisk mening är tillräckligt så finns det i många fall anledning att ifrågasätta ett sådant samtycke från ett etiskt perspektiv eftersom de som rutinmässigt skriver under ett samtycke för användning av en tjänst i många fall inte förstår vad de går med på att deras data används till. Lagen är och kommer aldrig vara fullt tillräcklig för att hantera problem som relaterar till personlig integritet, se även diskussionen i [44].

Ytterligare en slutsats kring de etiska diskussionerna kring praktisk användning av avanonymiseringstekniker för exempelvis polisiära syften är att det är viktigt att ta hänsyn till hur tekniken är tänkt att användas. Om det handlar om att försöka utesluta att en individ ligger bakom ett anonymt hot så skulle det exempelvis vara mer etiskt försvarbart än att försöka peka ut en definitiv gärningsman i en rättegång. På samma sätt spelar god kunskap om algoritmers träffsäkerhet och risken för felaktiga utpekanden i olika typer av situationer en viktig roll för att avgöra huruvida det är etiskt försvarbart att använda en viss teknik. Detta är ytterligare ett argument för varför forskning kring avanonymisering är viktig och i sin tur etiskt försvarbar.

Oavsett om det är forskningsmässig eller praktisk användning som övervägs så är det viktigt att försöka minimera intrång på den personliga integriteten. I forskningssyfte är det många gånger möjligt att göra experiment på syntetisk data eller data som gjorts tillgänglig för just forskning. I de fall det är möjligt så bör experiment utföras på denna typ av data. En begränsning med denna typ av studier är dock att det inte nödvändigtvis ger en rättvis bild av hur en algoritm skulle prestera under verkliga förhållanden, vilket kommenterats på flera olika ställen i

denna rapport. I de fall som experiment måste utföras på verklig data så bör experimenten utföras utifrån riktlinjer som minimerar intrång på den personliga integriteten, exempelvis genom att ingen människa läser de inlägg som används för att extrahera särdrag och att inga enskilda individers åsikter identifieras.

Vad gäller operativ användning så är slutsatserna utifrån ett etiskt perspektiv att det ytterst beror på vilket syftet med avanonymiseringen är. Kommersiell användning av avanonymisering är enligt författarna till denna rapport högst etiskt tveksamt och vad gäller användning av denna typer av tekniker för identifiering av politiska dissidenter, oliktänkande m.m. så kan det givetvis aldrig accepteras.

5 Slutsatser

Avanonymisering är ett samlingsnamn för en rad olika tekniker som på olika sätt kan nyttjas för att utvinna information om användaren bakom ett anonymt alias, innefattande såväl författarigenkänning, användarprofilering och aliasmatchning.

Författarigenkänningsalgoritmer baserade på stylometriska särdrag har visat sig vara robusta och i allmänhet erbjuda hög träffsäkerhet vid ett begränsat antal kandidatförfattare och tillgång till stora mängder data att bygga representativa klassificerare ifrån. På senare år har även studier presenterats som visar att vissa metoder skalar bra även till situationer där det finns flera tusen kandidatförfattare och det endast finns tillgång till en begränsad mängd inlägg, dock med en försämrad träffsäkerhet. Forskning kring att komplettera författarigenkänning med information om beteenden såsom tiden för skapande av en anonym individs inlägg har också visat sig ha stor potential som komplement till stylometri, givet tillräckliga mängder användargenererat innehåll. Det största praktiska problemet med författarigenkänning är att det i många fall är svårt att hitta representativ data för varje kandidatförfattare att träna upp algoritmerna på samt att det är svårt att skapa en uttömmande mängd av potentiella kandidatförfattare.

Algoritmer för användarprofilering är i många fall mindre väl utvärderade än deras motsvarigheter för författarigenkänning. De slutsatser som går att dra utifrån tillgängliga publicerade studier är att det är potentiellt möjligt att försöka härleda en anonym användares ålder och kön givet analys av användandet av pronomen, negationer och prepositioner, dock varierar träffsäkerheten med typen av domän. Även mer kontextberoende ord som har att göra med individens intressen kan säga mycket om den anonyma användarens kön och ålder. Även mer känsliga personuppgifter som politiska åsikter, sexuell läggning och personlighetstyp har i relativt nyligen utförda studier visat sig gå att få fram med oväntat bra träffsäkerhet baserat på användarens beteende på sociala medier, såsom vilka typer av nyheter som användaren aktivt ”gillar”. Även geolokalisering kan ses som en specifik typ av användarprofilering som kan användas för att med varierande träffsäkerhet försöka positionerna (anonyma) användare utifrån information såsom vilka platser och slanguttryck de nämner i sina inlägg, vilken typ av innehåll de länkar till samt känd geoposition för några av de användare de interagerar med i sociala medier.

För att länka samman användarprofiler som tillhör en och samma användare har flera olika typer av aliasmatchningsalgoritmer utvecklats. I de fall användaren inte aktivt undviker sammanlänkning har algoritmer baserat på valda användarnamn visat sig vara kraftfullt, men detta fungerar inte när användaren medvetet väljer väldigt olika aliasnamn för olika konton. I dessa fall har dock stylometriska och publiceringstidsbaserade särdrag visat god potential givet ett tillräckligt stort dataunderlag. Preliminära studier visar även att information utvunnen från de bilder som användare tar med sina mobilkameror och laddar

upp på sociala medier kan vara användbara för att länka samman användarens olika konton.

Den sammanställning av avanonymiseringstekniker som redogjorts för i denna rapport är såvitt författarna vet den mest heltäckande översikt över avanonymiseringstekniker för sociala medier som gjorts. Goda sammanställningar över litteratur inom författarigenkännings- och geolokaliseringsområdena har gjorts sedan tidigare men ingen som ger en holistisk översyn över flera sorters avanonymiseringstekniker.

Det finns i många situationer ett stort potentiellt värde i att använda avanonymiseringstekniker för exempelvis polisiära och försvarsunderrättelserelaterade syften. Vid sådan användning är det dock viktigt att inte applicera en teknik utan att först förstå dess underliggande antaganden och att analysera dess lämplighet för den aktuella problemställningen. En genomgång av forskningslitteraturen har visat på att den träffsäkerhet som rapporterats i många studier kan vara fullt tillräcklig för att vara praktiskt användbar, dock har det visats att vissa av metoderna inte klarar av att skala upp till mer realistiska förhållanden.

Sammanfattningsvis finns det tekniskt en stor potential för användning av denna typ av tekniker och de närmaste åren kan det förväntas ytterligare stora framsteg som ökar de tekniska möjligheterna till framgångsrik avanonymisering. Det svenska och europeiska juridiska landskapet för praktisk tillämpning av avanonymisering och andra typer av automatiserad analys av sociala medier är dock mer snårigt och osäkert än någonsin på grund av osäkerheter i vilka effekter den kommande EU-förordningen kommer att ha. För tillämpningar som har att göra med bevarandet av rikets säkerhet så är det dock högst rimligt att anta att syftet fortsatt kommer att väga tyngre än den enskilde individens rätt till personlig integritet. Vid sådan eventuell användning bör dock ändå etiska riktlinjer sättas upp för att minimera onödiga övertramp av svenska och utländska medborgares rätt till personlig integritet.

Referenser

[1] P. Juola, ”Authorship attribution,” *Foundations and Trends in Information Retrieval,* pp. 233-334, 2006.

[2] D. I. Holmes, ”Authorship attribution,” *Computers and the Humanities,* vol. 28, nr 2, pp. 87-106, 1994.

[3] T. C. Mendenhall, ”The characteristic curves of composition,” *Science,* vol. IX, pp. 237-249, 1887.

[4] B. Brainerd, Weighing Evidence in Language and Literature: A Statistical Approach, University of Toronto Press, 1974.

[5] H. Sichel, ”On a Distribution Representing Sentence-Length in Written Prose,” *Journal of the Royal Statistical Society,* vol. 137, pp. 25-34, 1974.

[6] H. H. Somers, ”Statistical Methods in Literary Analysis,” i *The Computer and Literary Style,* Ohio, Kent State University Press, 1966.

[7] F. Mosteller och D. L. Wallace, Inference and Disputed Authorship: The Federalist, Reading: Addison-Wesley, 1964.

[8] G. U. Yule, The Statistical Study of Literary Vocabulary, Cambridge University Press, 1944.

[9] A. Q. Morton, ”Once. A Test of Authorship Based on Words which are not Repeated in the Sample,” *Journal of the Association for Literary and Linguistic Computing,* vol. 1, nr 1, pp. 1-8, 1986.

[10] H. S. Sichel, ”Word Frequency Distributions and Type-Token Characteristics,” *Mathematical Scientist,* vol. 11, pp. 45-72, 1986.

[11] M. Colosimo, R. Graef, S. Lampert och M. Peterson, ”State of the art biometrics excellence roadmap,” US Department of Justice Federal Bureau of Investigation, 2009.

[12] A. Abbasi och H. Chen, ”A stylometric approach to identity-level identification and similarity detection in cyberspace,” *ACM Transactions on Information Systems,* vol. 26, nr 2, pp. 7:1-7:29, 2008.

[13] M. Koppel, J. Schler och S. Argamon, ” Authorship Attribution in the Wild,” *Language Resources and Evaluation,* vol. 45, nr 1, pp. 83-94, 2011.

[14] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin och D. Song, ”On the Feasibility of Internet-Scale Author Identification,” i *2012 IEEE Symposium on Security and Privacy*, 2012.

[15] J. R. Rao och P. Rohatgi, ”Can pseudonymity really guarantee privacy?,” i *Proceedings of the 9th USENIX Security Symposium*, Denver, 2000.

[16] F. Johansson, L. Kaati och A. Shrestha, ”Time Profiles for Identifying Users in Online Environments,” i *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, Haag, 2014.

[17] M. Koppel, S. Argamon och A. R. Shimoni, ”Automatically Categorizing Written Texts by Author Gender,” *Literary and Linguistic Computing*, vol. 17, nr 4, pp. 401-412, 2002.

[18] J. Schler, M. Koppel, S. Argamon och J. Pennebaker, ”Effects of Age and Gender on Blogging,” i *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006.

[19] N. Cheng, R. Chandramouli och K. Subbalakshmi, ”Author gender identification from text,” *Digital Investigation*, vol. 8, nr 1, pp. 78-88, 2011.

[20] S. Argamon, M. Koppel, J. W. Pennebaker och J. Schler, ”Automatically profiling the author of an anonymous text,” *Communications of ACM*, vol. 52, nr 2, pp. 119-123, 2009.

[21] M. Kosinski, D. Stillwell och T. Graepel, ”Private traits and attributes are predictable from digital records of human behavior,” i *Proceedings of the National Academy of Sciences (PNAS)*, 2013.

[22] W. Youyou, M. Kosinski och D. Stillwell, ”Computer-based personality judgments are more accurate than those made by humans,” i *Proceedings Of The National Academy Of Sciences (PNAS)*, 2015.

[23] J. Eisenstein, B. O'Connor, N. A. Smith och E. P. Xing, ”A Latent Variable Model for Geographic Lexical Variation,” i *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, 2010.

[24] S. Roller, M. Speriosu, S. Rallapalli, B. Wing och J. Baldridge, ”Supervised text-based geolocation using language models on an adaptive grid,” i *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Stroudsburg, 2012.

[25] Z. Cheng, J. Caverlee och K. Lee, ”You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users,” i *Proceedings of the 19th*

ACM International Conference on Information and Knowledge Management, New York, 2010.

[26] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey och M. Mühlhäuser, ”A Multi-Indicator Approach for Geolocalization of Tweets,” i *Proceedings of the 2013 International AAAI Conference on Weblogs and Social Media*, 2013.

[27] B. Hecht, L. Hong, B. Suh och E. H. Chi, ”Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles,” i *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[28] M. Dredze, M. J. Paul, S. Bergsma och H. Tran, ”Carmen: a Twitter geolocation system with applications to public health,” i *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI*, Bellevue, 2013.

[29] J. Mahmud, J. Nichols och C. Drews, ”Home location identification of Twitter users,” *ACM Transactions on Intelligent Systems and Technology,* vol. 5, nr 3, 2014.

[30] L. Backstrom, E. Sun och C. Marlow, ”Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity,” *Proceedings of the 19th International Conference on World Wide Web*, pp. 61-70, 2010.

[31] A. Sadilek, H. Kautz och J. P. Bigham, ”Finding Your Friends and Following Them to Where You Are,” i *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, Seattle, Washington, USA, 2012.

[32] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu och D. Ruths, ”Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice,” i *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 2015.

[33] R. Zafarani och H. Liu, ”Connecting Corresponding Identities across Communities,” i *Proceedings of the Third International Conference on Weblogs and Social Media*, 2009.

[34] D. Perito, C. Castelluccia, M. Kaafar och P. Manils, ”How unique and traceable are usernames?,” *Privacy Enhancing Technologies*, pp. 1-17, 2011.

[35] R. Zafarani och H. Liu, ”Connecting Users Across Social Media Sites: A Behavioral-modeling Approach,” i *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

[36] I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas och E. P. Markatos, ”Using social networks to harvest email addresses,” i *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*, 2010.

[37] S. Liu, S. Wang, F. Zhu, J. Zhang och R. Krishnan, ”HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling,” i *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014.

[38] F. Johansson, L. Kaati och A. Shrestha, ”Timeprints for identifying social media users with multiple aliases,” *Security Informatics* , vol. 4, nr 7, 2015.

[39] M. Ashcroft, F. Johansson, L. Kaati och A. Shrestha, ”Multi-Domain Alias Matching Using Machine Learning,” i *Proceedings of the third European Network Intelligence Conference*, Wroclaw, Poland, 2016.

[40] F. Johansson, L. Kaati och A. Shrestha, ”Detecting Multiple Aliases in Social Media,” i *Proceedings of the 2013 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Istanbul, Turkey, 2013.

[41] F. Bertini, R. Sharma, A. Iannì och D. Montesi, ”Smartphone verification and user profiles linking across social networks by camera fingerprinting,” i *Proceedings of the 7th EAI International Conference on Digital Forensics and Cyber Crime*, Seoul, South Korea, 2015.

[42] M. Brennan, S. Afroz och R. Greenstadt, ”Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity,” *ACM Transactions on Information and System Security,* vol. 15, nr 3, pp. 12:1-12:22, 2012.

[43] S. Afroz, M. Brennan och R. Greenstadt, ”Detecting Hoaxes, Frauds, and Deception in Writing Style Online,” i *2012 IEEE Symposium on Security and Privacy*, 2012.

[44] L. v. Wel och L. Royakkers, ”Ethical issues in web data mining,” *Ethics and Information Technology,* vol. 6, pp. 129-140, 2004.

FOI är en huvudsakligen uppdragsfinansierad myndighet under Försvarsdepartementet. Kärnverksamheten är forskning, metod- och teknikutveckling till nytta för försvar och säkerhet. Organisationen har cirka 1000 anställda varav ungefär 800 är forskare. Detta gör organisationen till Sveriges största forskningsinstitut. FOI ger kunderna tillgång till ledande expertis inom ett stort antal tillämpningsområden såsom säkerhetspolitiska studier och analyser inom försvar och säkerhet, bedömning av olika typer av hot, system för ledning och hantering av kriser, skydd mot och hantering av farliga ämnen, IT-säkerhet och nya sensorers möjligheter.

FOI
Totalförsvarets forskningsinstitut
164 90 Stockholm

Tel: 08-55 50 30 00
Fax: 08-55 50 31 00

www.foi.se