Big data, small data och etik i internetforskningen: utmaningar för samhällsvetenskapen

Gustafsson, Nils

Published in:
Samhällsvetenskapliga fakulteten i Lund - en vital 50-åring

2015

Big data, small data och etik i internetforskningen: utmaningar för samhällsvetenskapen

Nils Gustafsson

ABSTRACT

I denna text vill jag göra tre poänger: 1) internet är idag och kommer över överskådlig tid att vara ett oerhört viktigt fält för forskning och datainsamling inom samhällsvetenskapen; 2) det finns en rad olika problem – metodmässiga, ekonomiska och etiska – med datainhämtning i stor skala; 3) det etiska regelverket för denna typ av forskning är i Sverige outvecklat och måste ses över.

SOCIALA MEDIER OCH SOCIALA SMITTOR

Under en vecka i januari 2012 ändrade den sociala nätverkssajten Facebook nyhetsflödet för knappt 700 000 användare. Man utförde experiment i syftet att testa graden av så kallad emotionell smitta, dvs. hur mycket vi påverkas av de känslor våra vänner visar. Enkelt uttryckt utfördes experimenten så att i ett försök minskades mängden statusuppdateringar med positivt innehåll. I ett uppföljande försök minskades i stället mängden ”negativa” statusuppdateringar. Resultaten publicerades i en studie i PNAS[1] som visade att om mängden positiva ord minskades i flödet, minskade också försökspersonernas användning av positiva ord. Det samma gällde för det fall de negativa orden minskades – då minskade också försökspersonernas användning av negativa ord. Detta anfördes som ett belägg för att emotionell smitta existerar – vi påverkas av de känslor som folk omkring oss uttrycker, i alla fall på Facebook, och även om effekterna var små kunde detta ha betydelse för folkhälsan på aggregerad nivå, argumenterade författarna.[2] Åtminstone om man tror att de känslor folk ger uttryck för på Facebook kan säga något om vilka känslor de verkligen har.

Det var dock inte detta som debatten främst handlade om i spåren av publiceringen, utan huruvida det var etiskt försvarbart eller ens lagligt att

genomföra den här typen av studier. De Facebookanvändare som deltog i experimentet hade varken informerats eller givit sitt samtycke. Det fanns därför ingen möjlighet för användare att avstå från att delta. De informerades inte heller i efterhand om att de hade deltagit. Detta kritiserades häftigt under de efterföljande veckorna utifrån två utgångspunkter: dels att Facebook skulle ha ”gjort användarna ledsna”; dels att detta rörde sig om forskning som avser människor och därför borde underkastats en etisk prövning.[3] Två av författarna var verksamma vid Cornell University, som har ett etiskt regelverk för forskning som avser människor vilket kräver informerat samtycke. Författarna anförde dock att Cornell University inte ansåg att experimenten föll under det etiska regelverket eftersom det rörde sig om ett internt projekt på Facebook.[4] Dessutom hade de två universitetsknutna författarna inte samlat in data utan endast hjälpt till med att designa experimenten och skriva artikeln.[5] Projektet var förenligt med Facebooks dataanvändningspolicy, där användare bland annat godkänner att allt innehåll och alla data de lämnar ifrån sig får användas för forskning. Kravet på informerat samtycke ansågs därför vara uppfyllt.[6] Det anfördes dessutom att Facebook alltid manipulerar användarnas nyhetsflöde genom de algoritmer som bestämmer vilket innehåll som visas och i vilken ordning, baserat på vad Facebook tror att användarna vill se, och att man därför inte hade förändrat deltagarnas miljö på ett sätt som innebar fler risker än vad de normalt möter i vardagslivet.[7] Experimenten var alltså lagliga, men var de etiskt försvarbara? Det beror på vad man kan anse vara försvarbart vad gäller forskning på privatpersoners data på internet, och som skall visas i det följande är inte det helt uppenbart.

STORA DATA OCH STORA PROBLEM

När jag själv påbörjade mitt avhandlingsprojekt om sociala medier och politiskt deltagande 2007 kunde jag skaffa mig en hyfsad överblick över den samhällsvetenskapliga forskning som gjordes kring sociala medier som Facebook, men fältet exploderade snabbt och det är nu omöjligt att ha en fullständig överblick över utvecklingen inom ens specialintresse. Sociala medier har blivit en naturlig arena för samhällsvetenskaplig forskning och datainhämtning.

Detta handlar inte primärt om att sociala medier skulle vara ett nytt och spännande studieobjekt – nyhetens behag försvann för ett halvt decennium sedan – utan för att det ter sig så lockande lätt att där hitta material

för vilken aspekt av mänskligt beteende eller handlande man än är intresserad av.

De sociala mediernas datarikedomar och enkelheten med vilken man kan hämta ner stora mängder data och underkasta dessa data vetenskaplig analys är en del av den trend som pågått ett antal år men som växer sig stadigt allt starkare: big data.

Big data är ett samlingsnamn för en rad olika typer av datamaterial – offentliga register, sociala medier, information från sökmotorer och teleoperatörer – som har det gemensamt att de är stora, sökbara och möjliga att aggregera.[8] Det rör sig i allmänhet om redan existerande data som samlats in i olika syften och som kan användas för att studera något specifikt eller för att rota runt i största allmänhet för att försöka finna korrelationer (data mining). Ofta är det möjligt att använda anpassad mjukvara som möjliggör automatiserad datainhämtning, vilket ytterligare förenklar processen. Big data förknippas dessutom ibland med tanken att det bland dessa stora datamängder är möjligt att med en hög grad av noggrannhet och objektivitet ”en högre form av intelligens och kunskap”.[9] En vanlig kritik mot användning av big data är att den ofta är mer eller mindre teorilös. Chris Anderson, journalist och författare och bland annat känd för uttrycket ”the long tail”, uttryckte det så: ”With enough data, the numbers speak for themselves.”[10]

I en tid av svårigheter med att få rimliga svarsfrekvenser på vanliga enkätundersökningar är internet och de sociala medierna något av ett Schlaraffenland – här lägger människor frivilligt – åtminstone ibland – ut en massa information som bara finns där, ofta lätt tillgänglig. Med några enkla grafiska knep kan man få ut oerhört vackra visualiseringar, ofta i form av nätverkskartor av olika slag, som upplevs som intuitivt begripliga. Nästan i realtid kan man se mönster som uppträder i beteendet hos miljontals människor och upptäcka saker om oss själva vi inte hade kunnat finna ut på ett annat sätt.

Jag vill understryka att detta till stor del är viktigt och bra. Det finns dock en uppsättning problem med detta, som samtidigt har olika tänkbara lösningar. De utmaningar som big data ställer på samhällsvetenskapen kommer att bli allt mer tydliga under de kommande åren, och det är därför viktigt för alla samhällsvetare att reflektera över dessa frågor. Jag kommer i det följande främst att hämta mina exempel från sociala nätverkssajter, men resonemangen är tillämpliga på en rad olika typer av data.

INFORMATIONSMONOPOLISTERNA OCH DE DATARIKA FORSKARNA

Det finns flera metodproblem med att använda datamängder från sociala medier. Flera av de problemen härstammar från det faktum att en allt större mängd av mänsklighetens kommunikation och information samlas hos ett fåtal kommersiella aktörer.[11] Dessa informationsmonopolister kontrollerar hur kommunikationen och informationen samlas in, sparas och visas för användarna. Facebooks algoritmer för att manipulera nyhetsflödet är ett exempel på detta.

Innehåll från Twitter är jämförelsevis lätt att komma åt – Twitter har ett så kallat öppet API vilket möjliggör att man direkt eller med hjälp av olika typer av verktyg lätt kan ladda ner och analysera tweets utifrån exempelvis vissa söktermer, så kallade hashtags och angivna tidsramar. Det finns dock begränsningar i detta, exempelvis vad gäller hur många tweets man kan få tillgång till och hur många söktermer man kan använda osv. Det är betydligt lättare att hämta in tweets i realtid än att göra historiska sökningar. Några av de begränsningar som finns är möjliga att komma runt med hjälp av kommersiella aktörer som säljer vidare innehåll från Twitter, eller (naturligtvis) genom att samarbeta med företaget Twitter självt.[12] Slutligen finns det information som är otillgänglig, som privata direktmeddelanden.

Även om det alltså finns problem med att komma åt Twitter är det betydligt svårare att komma åt data från Facebook. Medan tweets är offentliga i den meningen att man inte behöver vara ”vän” med användaren eller ens själv vara registrerad användare av Twitter för att kunna läsa tweets, är detta ett krav för att komma åt statusuppdateringar från Facebookprofiler. Därigenom tolkas Facebookprofiler som privata eller halvprivata, och många studier har därför begränsat sig till att studera så kallade Pages, som är Facebooksidor vem som helst kan läsa.[13] Detta leder dock till att en överväldigande mängd data helt enkelt är onåbara för de flesta forskare.

Det har dock gjorts flera uppmärksammade studier där premissen har varit obegränsad tillgång till information från Facebookanvändare och möjlighet att göra experiment genom olika manipulationer.[14] Förutsättningen har i de fallen varit ett intimt samarbete mellan forskare knutna till universitet och Facebooks egen forskningsavdelning.

Eftersom Facebook är världens största sociala nätverk med enligt egen utsago över en miljard aktiva användare[15] leder detta till att en av mänsklighetens viktigaste källor till information och kommunikation i princip är otillgänglig för oberoende forskning. De forskare som får den priviligierade positionen att åtnjuta direkttillgång till data kan vara förhindrade att

dela med sig av exakt hur data har tagits fram och vilka algoritmer som har använts, eftersom detta kan anses vara företagshemligheter.[16] Detta leder till en klyfta mellan datarika och datafattiga forskare, som på sikt kan bli mycket problematisk för samhällsvetenskapen.

ETISK MEDVETENHET OCH ETISKA REGELVERK

Enligt den svenska lagen (2003: 460) om etikprövning som avser människor skall regionala etikprövningsnämnder bedöma forskning som avser bland annat ”känsliga personuppgifter” (det finns en rad andra kriterier, men de är främst tillämpliga på medicinsk forskning). Med känsliga personuppgifter avses bland annat politiska åsikter, vilket inte minst berör min egen forskning. Sådan forskning som avses i lagen får inte utföras om den inte har prövats och godkänts av en etikprövningsnämnd. Forskning ska utföras ”med respekt för människovärdet” och om ”de risker som den kan medföra för forskningspersoners hälsa, säkerhet och personliga integritet uppvägs av dess vetenskapliga värde”. I normalfallet skall även information ges till forskningspersoner och samtycke inhämtas. Det finns ett undantag för insamling av känsliga personuppgifter i den svenska Personuppgiftslagen (1998: 224) som tillåter att detta kan ske utan informerat samtycke om ”det är omöjligt eller innebär en oproportionerligt stor arbetsinsats”.

Utöver den lagliga regleringen finns det även ett stort antal riktlinjer och råd, bland annat Vetenskapsrådets ”God forskningssed”. Denna rapport anger att man i de fall då man gör så kallade dolda observationer, där alltså de medverkande inte har informerats i förväg, måste informera dessa i efterhand. Detta ter sig svårt att göra i de fall då t ex Twitters öppna API används för att studera enskilda personers inlägg. Å andra sidan präglas Vetenskapsrådets text av att internetmetoder inte nämns överhuvudtaget. De observationsmetoder som anges som exempel handlar om t ex dold videoinspelning och inte om att ladda ner texter på nätet.

Vad gäller forskning som utförs på internet finns inga bindande regler. Vetenskapsrådet – som är den svenska myndighet som har ett övergripande ansvar för forskningsetiska frågor – hänvisar i stället till IT-kommissionens ”God etik på nätet” från 1998 (!), samtidigt som man hävdar att ”Forskning med hjälp av Internet är ännu i sin linda” (!!) AOIR:s ”Ethical decision-making and Internet research” och Den nasjonale forskningsetiske komité for samfunnsvitenskap og humanioras ”Forskningsetiske retningslinjer for forskning på Internett” från 2003

(som skall uppdateras 2015).[17] Det är således uppenbart att det behövs en förnyelse av riktlinjerna i den svenska kontexten.

I min egen praktiska verksamhet har jag lagt märke till att det inte tycks finnas en hög nivå av medvetenhet om de forskningsetiska principer och regelverk som finns. Samhällsvetenskaplig forskning studerar ofta enskilda personer och använder vad som kan anses vara känsliga personuppgifter. Trots detta är det sällan som man lämnar in ansökningar till etikprövningsnämnderna, om inte detta är ett uttalat krav från anslagsgivare. Naturligtvis gäller även att forskning som inte kan anses beröra känsliga personuppgifter måste prövas etiskt, om inte annat så av forskaren själv.

Sammantaget verkar det alltså som om man i det svenska fallet i stor utsträckning får lita till allmänna etiska forskningsprinciper och sin egen bedömningsförmåga.

Ett problem som jag ser det är hur känsliga personuppgifter kan anses vara om personer frivilligt offentliggör dem på nätet. Detta hänger samman med en allmän diskussion om integritet och var gränsen för privatlivet går i en tid då denna gräns verkar förskjutas. Ovan har jag nämnt synen på Twitter som ett offentligt forum och Facebook som ett halvprivat forum.

Problemet med detta att tweets – utan att användaren tillfrågas eller ens informeras – approprieras för syften och sammanhang som användaren inte har kunnat förutse. Hur bör vi hantera exempelvis de tweets som skrevs av ungdomar på Utøya? Eller för den delen blogginlägg som skrivs av tonåringar med självskadebeteende. Här krävs personligt omdöme, men också en diskussion om var gränserna ska dras. Det är inte rimligt att betrakta allt tillgängligt material på nätet på samma sätt som en publicerad text i en dagstidning. Jag menar att det är rimligt att i större utsträckning inhämta godkännande vid användning av data från individer.

Att det finns en gråzon och en osäkerhet kring detta illustreras inte minst av det pågående lagstiftningsarbete som finns kring hanteringen av registerdata. 2012 beslutade Datainspektionen att Göteborgs universitet skulle ”upphöra med insamling och övrig behandling av personuppgifter i material insamlat inom ramen för svensk nationell datatjänst”.[18] Detta eftersom data som deponerats hos SND skulle kunna användas för ny forskning som inte i förväg definierats. Händelsen kan ses som illustrativ för hanteringen av integritetskänsliga personuppgifter inom samhällsvetenskaplig forskning.

2014 föreslog regeringens utredare Bengt Westerberg att SND skulle bli en svensk arkivmyndighet och att flera ändringar i bl.a. Etikprövningslagen och Offentlighets- och sekretesslagen skulle göras för att underlätta

användningen av registerdata med hänsyn till lagringen av personuppgifter.[19] Utredningen hänvisade intressant nog till problemet med att folk frivilligt lägger upp information på exempelvis Facebook som sedan kan spridas och användas i andra sammanhang, men kopplade inte direkt ihop detta med forskning. Men, som utredningen anför: ”Att den tekniska utvecklingen rymmer stora risker för den personliga integriteten är naturligtvis inget argument för minskat integritetsskydd vid forskning. ”[20]

EN UTMANING TILL OSS SJÄLVA

Det inledande exemplet med manipulationen av Facebookflöden utgör i själva verket ett bra exempel på alla de utmaningar som jag tagit upp här. Problemet med bias i data och svag teori (hur vet vi att statusuppdateringar på sociala nätverkssajter är indikationer på författarnas känslor?); problemet med informationsmonopolister och datarika forskare (Facebook själv genomförde projektet under förutsättningar som vanliga forskare inte har, och det finns så att säga bara ett Facebook); och det saknas välfungerande etiska regelverk och en etisk medvetenhet som skyddar individers privata data, inte minst i Sverige.

Vi bör som samhällsvetare reflektera över dessa problem, anstränga oss för att hitta lösningar på dem, och vara aktiva i den allmänna debatten för att på sätt bidra till en ansvarsfull och etiskt försvarbar samhällsvetenskaplig forskning på big data som dessutom ger trovärdiga resultat.

Som forskare bör vi enligt min mening

a. sträva efter att uppnå idealet om informerat samtycke även när vi behandlar stora datamängder som i princip är allmänt tillgängliga – om de är ”dolda” (Facebookprofiler) eller ”öppna” (Twitter) *borde inte spela någon roll, och inte heller graden av anonymisering*

b. detta gäller självfallet även kvalitativa metoder, och även när det är offentliga personer som studeras med hjälp av nätmetoder

c. övriga etiska principer om att undvika skada osv. bör alltid gälla

d. forskare som samarbetar med informationsmonopolister och andra kommersiella och statliga aktörer måste hålla sig till etiska principer

Dessutom bör Vetenskapsrådet ta initiativ till ett utarbetande av ett etiskt regelverk för internetforskning.

Noter

1. Kramer et al, 2014.
2. ibid, 8790.
3. Meyer, 2014.
4. Verma, 2014.
5. Meyer, 2014.
6. Facebook har överhuvudtaget vida möjligheter att utnyttja användarnas data (Facebook 2014a). Det sprids med jämna mellanrum rykten om att man genom att posta en viss text med juridisk klang kan hindra Facebook från att använda data hur de vill, men det går inte i efterhand att ta tillbaka de medgivanden man gjorde när man skaffade ett användarkonto (se t ex Newman 2014).
7. ibid.
8. Boyd & Crawford 2012, 663.
9. ibid, min övers)
10. Anderson, 2008.
11. Gustafsson, 2013, 40; jfr Hindman, 2008. Jag väljer här helt att bortse från en rad fundamentala metodproblem som finns rörande användningen av innehållsdata från sociala nätverkssajter i samhällsvetenskapliga studier. En sammanfattning av problem som rör snedvridna populationer, förekomsten av ickemänskliga konton (botar) och svårigheten att dra slutsatser om mänskligt beteende utifrån innehåll i sociala medier finns bland annat i Ruths och Pfeffer (2014). För ett resonemang om att användning av sociala medier bör studeras genom traditionella intervjuer och enkäter eftersom innehållsanalyser missar osynligt innehåll, se Gustafsson, 2013, 54ff).
12. Bruns och Liang, 2012.
13. Se t ex Larsson och Kalnes, 2014.
14. T ex den ovan refererade Kramer et al, 2014; Bond et al, 2012.
15. Facebook, 2014b.
16. Ruths och Pfeffer, 2014.
17. Vetenskapsrådet, 2013; AOIR, 2012; NESH, 2003/2014.
18. Datainspektionen, 2012.
19. SOU 2014:45.
20. ibid, 247.

Referenser

AOIR. 2012. *Ethical decision-making and Internet research*. http://aoir.org/reports/ethics2.pdf (Hämtad 2014-11-05).

ANDERSON, CHRIS. 2008. "The end of theory: will the data deluge makes the scientific method obsolete?". *The Wire,* 23 juni 2008. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (Hämtad 2014-12-02).

BOND, ROBERT M., FARISS, CHRISTOPHER J., JONES, JASON J., KRAMER, ADAM D. I., MARLOW, CAMERON, SETTLE, JAIME E. & JAMES H. FOWLER, 2012. "A 61-million-person experiment in social influence and political mobilization". *Nature* 489, 295–298.

BOYD, DANAH OCH CRAWFORD, KATE, 2012. "Critical questions for big data". *Information, Communication & Society* vol 15:5, 662–679.

NESH, 2003/2014. "Etiske retningslinjer for forskning på Internett. https://www.etikkom.no/forskningsetiske-retningslinjer/Samfunnsvitenskap-jus-og-humaniora/Internett-forsking/ (Hämtad 2014-11-05).

BRUNS, AXEL & LIANG, EUGENE L., 2012. "Tools and methods for capturing Twitter data during natural disasters," *First Monday* vol 17:4. http://firstmonday.org/article/view/3937/3193, (Hämtad 2014-12-02).

Datainspektionen, 2012. "Tillsyn enligt personuppgiftslagen (1998:204) – av Göteborgs universitet (Svensk nationell datatjänst)", dnr 811-2011. http://www.datainspektionen.se/Documents/beslut/2012-04-23-snd.pdf (Hämtad 2014-12-02).

IT-kommissionen, 1998. *God etik på nätet.* Rapport i samband med hearing i Riksdagshuset 1998-05-25. http://www.codex.vr.se/texts/God%20etik.pdf (Hämtad 2014-11-05).

Facebook, 2014a. "Policy för dataanvändning". https://www.facebook.com/about/privacy/ (Hämtad 2014-12-02).

Facebook, 2014b. "Company Info". http://newsroom.fb.com/company-info/ (Hämtad 2014-12-02).

GUSTAFSSON, NILS, 2013. *Leetocracy. Political participation, social network sites and inequality*. Lund: Lunds universitet, Statsvetenskapliga institutionen.

HINDMAN, MATTHEW, 2008. *The Myth of Digital Democracy*. Princeton: Princeton University Press.

KRAMER, ADAM D. I., GUILLORY, JAMIE T., & HANCOCK, JEFFREY T., 2014. "Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* vol 111:24, 8788-8790.

LARSSON, ANDERS & KALNES, BENTE, 2014. "'Of course we are on Facebook': Use and non-use of social media among Swedish and Norwegian politicians". *European Journal of Communication* vol 29:6, 653–667.

Lag (2003: 460) om etikprövning av forskning som avser människor.

MEYER, MICHELLE, 2014. "Everything you need to know about Facebook's controversial emotion experiment". *Wired* 30 juni 2014. http://www.wired.com/2014/06/everything-you-need-to-know-about-facebooks-manipulative-experiment/ (Hämtad 2014-12-02).

NEWMAN, LILY HAY, 2014. "Posting that viral Facebook copyright notice won't protect your data and never has". *Slate* 1 december 2014. http://www.slate.com/blogs/future_tense/2014/12/01/sharing_a_copyright_notice_doesn_t_exempt_users_from_facebook_s_data_use.html (Hämtad 2014-12-02).

Personuppgiftslagen (1998: 224).

RUTHS, DEREK & PFEFFER, JÜRGEN, 2014. "Social media for large studies of behavior". *Science* 346(6213): 1063–1064.

SOU 2014: 45. "Unik kunskap genom registerforskning". Betänkande av Regietsrforskningsutredningen.

VERMA, INDER, 2014. "Editorial expression of concern and correction". www.pnas.org/cgi/doi/10.1073/pnas.1412469111 (Hämtad 2014-12-02).

Vetenskapsrådet, 2011. *God forskningssed*. Vetenskapsrådets rapportserie 2011:1.

Vetenskapsrådet, 2013. http://www.codex.vr.se/forskninghumsam.shtml (Hämtad 2014-11-05).