Digitální stopa: Konec empirické sociologie?*

JAKUB SEDLÁČEK**
Fakulta sociálních věd / Filozofická fakulta, Univerzita Karlova, Praha

Digital Trace Data: The End of Empirical Sociology?

Abstract: In the 20th century empirical sociology possessed innovative methodological resources that granted it fairly exclusive access to understanding human social life. However, with the advent of digital technologies and increasing migration into the online world, this privilege started to shift into the hands of commercial entities. People of the 21st century now generate data with every step they take (both physical and virtual), and most of the current internet business models are built on the collection, analysis, and commercial utilisation of such data. The 'Digital Trace Data' left behind by billions of online users present an unprecedented opportunity for the study of their behaviour, characteristics, and social interactions. This article seeks to introduce readers to the world of Digital Trace Data and the three main areas in which such data are used: research, commerce, and surveillance. Examples of all three are given to illustrate the potential strengths, weaknesses, and associated risks. The article also seeks to provide warning of a future in which the largest repository of sociological data in history ends up locked behind the doors of commercial enterprises and government institutions.

Keywords: digital trace data, big social data, big data, computational social science, social media

Sociologický časopis / Czech Sociological Review, 2020, Vol. 56, No. 4: 471–490
https://doi.org/10.13060/csr.2020.015

Úvod

Autoři článku „The Coming Crisis of Empirical Sociology" už v roce 2007 upozorňovali, že zatímco ve 20. století sociologové zavedli do praxe mnohé inovativní metody (dotazníková šetření na reprezentativním vzorku populace, hloubkové rozhovory atd.), díky kterým se mohli pyšnit svým způsobem exkluzivním přístupem k porozumění společenskému životu, ve 21. století je a jejich metody co do

* Tento výstup vznikl v rámci projektu Grantové agentury Univerzity Karlovy (1084218) „Celebrita nových médií: Proměna světa celebrit v éře sociálních sítí " a projektu Specifického vysokoškolského výzkumu (260 596).

** Veškerou korespondenci posílejte na adresu: Mgr. Jakub Sedláček, Institut sociologických studií, Fakulta sociálních věd Univerzity Karlovy, Smetanovo nábř. 6, 110 01 Praha 1, e-mail: info@jakubsedlacek.cz.

© Sociologický ústav AV ČR, v. v. i., Praha 2020

schopnosti generovat a využívat sociální data nepřehlédnutelně předhánějí moderní informační technologie a jejich komerční provozovatelé [Savage, Burrows 2007: 885–886]. Systematický sběr dat je dnes jakýmsi samozřejmým, základním nastavením digitálních služeb, pevně zakotveným v podmínkách jejich užívání. Až konkrétní opatření typu legislativy GDPR se zaměřují na jeho omezení a dávají uživateli alespoň základní možnost s některými jeho formami nesouhlasit. Účast na všudypřítomném sběru dat není „opt-in" (dobrovolná, s vědomým souhlasem), ale „opt-out" (automatická, s teoretickou možností explicitního nesouhlasu). Nesouhlas se sběrem dat v celém rozsahu ale třeba v případě největší online sociální sítě Facebook znamená danou službu nepoužívat (což je samo o sobě v mnoha oblastech života problematické) a dále pomocí speciálního softwaru aktivně blokovat její nástroje zahrnuté v kódu na první pohled zcela nesouvisejících stránek a aplikací [Solon 2017]. Dokonce i v případě, že jsme Facebook nikdy nepoužívali, je možné, že nám v jeho databázi patří přinejmenším určitý „stínový" marketingový profil [Hill 2018], nad kterým paradoxně nemáme žádnou kontrolu právě proto, že Facebook nepoužíváme. Chceme-li používat internet, mobilní telefon nebo třeba bankovní účet a platební kartu, uniknout sběru dat je už nemožné.

Vzhledem k tomu, že internet se z místa setkávání několika technologických nadšenců postupně proměnil v prostor, kde žijí životy běžní lidé, je rozsah těchto dat skutečně nebývalý. Facebook dnes po celém světě využívá každý den 1,5 miliardy lidí, Instagram 500 milionů, WhatsApp dalších 500 milionů[1] [Statista 2019]. „Digitální stopa" představuje pomyslný soubor všech dat, která v systémech těchto služeb zůstávají jako otisk každodenních aktivit každého z nich. V současnosti se jedná primárně o data (a příslušná metadata) vzniklá interakcí v online prostředí: dotazy ve vyhledávačích, profilové informace, propojení v rámci sociální sítě, sdílené texty, fotografie a videa, soukromé zprávy a emaily atd. S rozšiřováním nositelné elektroniky a internetu věcí (IoT) však mohou data mít čím dál častěji původ i ve fyzickém světě: historie GPS polohy uživatele, jeho fyzická aktivita a zdravotní ukazatele, data z palubních systémů „chytrých" automobilů, systémů „chytré domácnosti", jako jsou vysavače, chladničky apod.

Následující dva oddíly přinášejí úvodní popis kontextu v podobě komerčního a „dohledového" využití digitální stopy a diskusi o jeho společenských dopadech. Představíme si největší medializované kauzy související s tím, jak velké online platformy (a složky státního aparátu) nakládají s daty. Ty mají totiž zcela zásadní vliv na povědomí a názor veřejnosti a v důsledku i na příslušnou legislativu, podmínky užívání služeb a dostupnost dat pro výzkum. Po oddílu, který nový typ dat (a jeho výhody i nevýhody) výzkumníkovi blíže představí, se právě problematice skokového zhoršení dostupnosti digitální stopy po aféře Cambridge Analytica a nástupu GDPR budeme věnovat. Zde spočívá i jedno z těžišť článku. Je jím apel na rozvoj akademického zájmu o studium (a boj o vý-

[1] Všechny tři uvedené služby navíc spadají pod stejného provozovatele, který sebraná data ve svých produktech kombinuje.

zkumnou dostupnost) digitální stopy. Ta již nyní představuje pravděpodobně největší repozitář sociologických dat v historii a je nemyslitelné ponechat ji v budoucnosti za zavřenými dveřmi komerčních subjektů, které do jejího mnohokrát prolustrovaného výseku jednou za čas velkomyslně dovolí vybranému výzkumníkovi nahlédnout. Závěrečný oddíl přináší přehled vybraných studií, které ilustrují potenciál konkrétních, dříve v článku naznačených využití digitální stopy. Základním cílem textu je představit čtenáři způsoby jejího praktického využití, přesto se v žádném případě nejedná o vyčerpávající metodologickou příručku. V tomto směru jde spíše o základní, motivační uvedení do oblasti, která disponuje potenciálem nahradit (nebo přinejmenším doplnit) mnohé tradiční metody empirické sociologie. Podaří-li se textu zaujmout, nechť čtenář využije seznam odkazované literatury jako výchozí bod k dalšímu studiu a metodologicky se inspiruje u předkládaných studií.

Komerční využití digitální stopy

Obchodní model většiny internetových služeb „zdarma" v současnosti stojí na přesném cílení obsahu a obchodních nabídek, na jehož začátku stojí právě nikdy nekončící analýza osobních dat uživatelů. O některých konkrétních metodách (například o analýze uživatelských konverzací a celého spektra dalších představitelných akcí v prostředí sociální sítě) se dočteme například v patentu Facebooku US20160283485A1 [Nowak, Eckles 2016]. Občané EU získali jistou (a ve srovnání se zbytkem světa nadprůměrnou) kontrolu nad využitím svých údajů až relativně nedávno díky nástupu GDPR.[2] V případě Facebooku přibyla možnost zakázat službě používat tzv. „sociální akce" v reklamách[3] nebo využívat k cílení reklamy údaje získané od třetích stran a naopak cílit reklamu na webech třetích stran skrz uživatelovy údaje z Facebooku. Nezměněné však zůstalo využití zájmového profilu uživatele vytvořeného z jeho chování v prostředí produktů společnosti Facebook, stejně jako jeho demografických údajů – s obojím uživatel souhlasí samotným používáním služby, byť dnes disponuje alespoň možností tento zájmový profil manuálně upravit. Rozhraní obrazovky s nastaveními GDPR souhlasů ovšem uživatele jasně vede k jednoduchému udělení všech potřebných povolení, zatímco změna nastavení v neprospěch provozovatele je viditelně upozaděna a vyžaduje od uživatele snažení navíc [Constine 2018].

Oproti době před online sociálními sítěmi dnes lidé o sobě veřejně sdílejí velké množství informací, často aniž by si uvědomovali, že i z dat zdánlivě zcela

[2] Nařízení Evropského parlamentu a Rady (EU) č. 2016/679 ze dne 27. dubna 2016, o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (obecné nařízení o ochraně osobních údajů).

[3] Reklama zobrazená některému z přátel uživatele mohla v takovém případě být uvozená třeba větou „Janu Novákovi se líbí ***".

nevinných je možné strojově odvodit údaje, které sami považují za citlivé. Crandall et al. [2010] například s pomocí místa a času pořízení 38 milionů fotografií ze sociální sítě Flickr zjistili, že je možné velmi úspěšně předpovědět existenci společenských vazeb čistě na základě toho, že jsou dva lidé opakovaně ve stejnou dobu na stejném místě. Ač se z pohledu sociologa nejedná o zjištění nijak šokující, snadná dostupnost tohoto typu dat o velkých populacích může mít podstatné dopady na soukromí. Na rizika veřejného oznamování vlastní polohy cíleně upozorňovala například stránka „vykraďte mě, prosím" (pleaserobme.com), která propojovala veřejně dostupná data ze sociálních sítí Twitter a Foursquare a následně celý internet informovala, že je uživatel mimo domov, a tedy připravený k vykradení [Quercia et al. 2011]. Studie postavené na datasetu aplikace myPersonality (kterou si blíže představíme v dalších oddílech) podobně úspěšně ukazují, že i kdyby uživatel nechtěl veřejně sdílet informace, jako je politické přesvědčení, sexuální orientace nebo vlastní psychologický profil, i nevinná digitální stopa v podobě seznamu sledovaných stránek na Facebooku může s dostatečně silným prediktivním modelem snadno posloužit k jejich „dopočítání".

Pravděpodobně nejznámější aféra související se zneužitím soukromých dat z online sociálních sítí se odehrála v souvislosti s poradenskou společností Cambridge Analytica. Ta se nejprve pokoušela koupit data projektu myPersonality, ale protože výzkumníci s komerčním použitím nesouhlasili [Kosinski 2018], došlo na zkopírování jejich metodiky a tvorbu aplikace vlastní. Ta v přímém srovnání nebyla zdaleka tak úspěšná jako její vzor (osobnostní dotazník vyplnilo „pouze" 300 tisíc lidí oproti šesti milionům u aplikace původní), zato ale využila v té době značně benevolentního nastavení samotného Facebooku, které až do roku 2015 umožňovalo získat digitální stopu nejen dotazníkových respondentů (těch, kteří aplikaci sběr dat explicitně povolili), ale i všech jejich facebookových přátel. Odtud pochází medializované číslo 87 milionů zcizených profilů. Sběr dat proběhl legální cestou a navzdory tvrzení četných titulků nešlo o prolomení zabezpečení Facebooku. Podmínky byly porušeny až komerčním využitím dat, k němuž autor aplikace Aleksandr Kogan povolení neměl [Cadwalladr, Graham-Harrison 2018]. Tři sta tisíc dotazníkových a 87 milionů z digitální stopy pomocí prediktivního modelu vytvořených osobnostních profilů potom mělo být použito k přesnému cílení reklamy v rámci politických kampaní, na nichž se Cambridge Analytica podílela. Nakolik cílení pomocí psychometrických profilů třeba americké prezidentské volby v roce 2016 skutečně ovlivnilo, zůstává nejasné. Stejně tak jestli vůbec mohlo dosáhnout většího efektu než zcela oficiální systém cílení reklamy na Facebooku, jehož role ve zmíněné kampani je nezpochybnitelná a zdaleka se neomezuje jen na takzvaný „ruský vliv" ve zvolení Donalda Trumpa, tedy cca 3 500 reklam ruské agentury zaměřených na „destabilizaci západní demokracie cílením na krajní názorové skupiny" [Penzenstadler, Heath, Guynn 2018].

Navzdory tvrzením velkých technologických firem nepředstavují nebezpečí pro soukromí zdaleka jen zákulisní aktivity hackerů (využívajících děr v zabezpečení) a na hraně zákona se pohybujících organizací typu Cambridge Analytica.

Značnou pozornost si zaslouží i rizika sběru dat v podání samotných provozovatelů služeb, kterým je zpřístupňujeme dobrovolně. Ať už se jedná o sběr informací o nákupním chování skrz zákaznické karty obchodů, přesně kategorizovanou historii pohybů na účtech v bance nebo výše zmíněné profilování za účelem cílení reklamy v podání většiny internetových gigantů: i zde se skrývá nebývalý potenciál v podobě kombinace nejrozličnějších zdrojů dat, které dohromady mohou mít mnohem větší výpovědní hodnotu, než bylo při souhlasu s jejich využitím zamýšleno. Vzpomeňme některé z mnoha medializovaných kauz Facebooku: například když se snažil získat zdravotní dokumentaci pacientů nemocnic [Kastrenakes 2018], případně právě data o pohybech na bankovních účtech a nákupech platebními kartami [Glazer, Seetharaman, Andriotis 2018], obojí pochopitelně za účelem propojení se svým stávajícím datasetem.

Bylo by krátkozraké tvářit se, že takováto propojení velkých, heterogenních datasetů nemohou být zdrojem pro vědu a společnost přínosného poznání, nebo dokonce přímo životy zachraňujících praktických nástrojů – v případě zdravotních dat se nutně nabízí třeba myšlenka využití umělé inteligence k včasné detekci dnes zcela neznámých a běžným okem nepostřehnutelných projevů nemocí ukrytých hluboko v digitální stopě uživatele a jeho následné vyzvání k návštěvě lékaře [Matz et al. 2017]. Potenciál kombinace big data a strojového učení je v tomto směru nedozírný. Představme si, že například jako výzkumník zabývající se prevencí civilizačních onemocnění v nedaleké budoucnosti máme k dispozici v minutových intervalech po mnoho let zaznamenávaný vývoj zdravotních ukazatelů[4] a fyzických aktivit milionů pacientů a k nim například kompletní GPS záznamy o jejich pohybu, transakce platební kartou mapující jejich stravování, kompletní přehled vazeb v rámci jejich sociální sítě nebo přesné údaje o kulturní a mediální spotřebě a texty veškerých jejich písemných komunikací. To vše jsou data, která díky nositelné elektronice (tzv. „wearables"), online sociálním sítím a skutečnosti, že takřka každý provozovatel služby sbírá co možná největší množství dat o svých zákaznících, existují již dnes a jejich množství i přesnost každým rokem narůstají. Zatímco v minulosti data v takovém množství nemělo smysl ani sbírat, neboť neexistovaly praktické cesty k jejich zpracování, strojové učení dnes v obřích datových souborech dokáže odhalit pravidelnosti mnohdy zcela mimo rozlišovací schopnosti lidského pozorovatele. Zároveň by ale bylo naivní svěřit tato data slepě do rukou subjektů a zákulisních hráčů, jejichž obchodní model stojí na neprůhledném a v mnoha ohledech nekontrolovaném komerčním využití dosaženého poznání.

[4] Již dnešní (stále vcelku primitivní) přístroje zvládají měřit faktory jako srdeční tep, krevní tlak, EKG, okysličení krve, tělesná teplota, podkožní tuk a jejich záběr i přesnost se neustále zvyšují.

Systémy dohledu

Digitální stopa nachází bohaté využití i jako nástroj dohledu (surveillance) ať už přímo v rukou totalitních režimů, nebo „pouze" složek státního aparátu a tajných služeb demokratických zemí. Například systém „sociálního kreditu", který se postupně implementuje v Číně, staví na technikách a technologiích známých ze Silicon Valley, ale vzhledem k často přímému napojení čínských technologických firem na stát mu nestojí prakticky nic v cestě. Cílem systému je s pomocí strojového učení analyzovat nejen online aktivitu jedince, ale i stále narůstající množství digitálně kvantifikovatelných aktivit ve světě offline. Výsledné „skóre občana" je postavené na množství faktorů, které nejsou vždy ani oficiálně známé. Patří tam například placení daní, pokuty, šíření „nevhodných" informací, užívání drog, cvičení, studium ideologických materiálů prostřednictvím příslušné aplikace Xuexi Qiangguo [Liao 2019] atd. Jedince má systém počínaje rokem 2020 povinně provázet každodenním životem a přinášet mu zvýhodnění (prominentní umístění profilu na seznamce, slevy na poplatcích za energie, půjčování věcí bez nutnosti složit zálohu, lepší úrokové sazby v bankách atd.) nebo naopak omezení (zákaz nákupu letenek nebo jízdenek na vlak [Kuo 2019], omezení rychlosti internetu, nemožnost ucházet se o vyšší posty ve státních firmách, nemožnost ubytování v lepších hotelích, neumožnění univerzitního studia dětí) [Ma 2018].

Chceme-li v dystopických vizích světa velkých dat a strojového učení zajít ještě o něco dál, nabízí se možnosti jako rozpoznávání emocí v reálném čase na základě fyziologických projevů zaznamenaných kupříkladu chytrými hodinkami [příklad viz Rincon et al. 2016]. Takový systém by byl teoreticky schopný rozpoznat například hněv, znechucení a další emoce při kontaktu třeba právě s ideologickým materiálem a následně v systému označit potenciálně rizikové jedince. Cesta k orwellovskému „ideozločinu" a příslušné „ideopolicii" je tak načrtnuta zcela jasně a v jisté míře technicky realizovatelná již dnes.[5]

Zapojení digitální stopy do systémů dohledu můžeme být svědky i mimo totalitní režimy tam, kde je potřeba bezpečnosti upřednostněna před svobodou a soukromím: Americký pohraniční úředník, který si od zájemce o vstup do země krom pasu vyžádá i účty na sociálních sítích [Larson 2017] nebo Americká národní bezpečnostní agentura (NSA), sbírající metadata telefonních hovorů [Landau 2013], odposlouchávající hovory a analyzující platby kartami, atd. Zařadit sem můžeme i projekty typu EET, jejichž základním cílem je zabránit daňovým únikům na straně obchodníků, ale které v momentě propojení například s databázemi provozovatelů platebních karet mohou poskytnout dokonalý přehled o spotřebním chování obyvatelstva v reálném čase. Méně dystopickou kategorií jsou systémy s dobrovolnou účastí, kam můžeme zařadit i aktuálně medializovaný český příklad využití digitální stopy v rámci tzv. „chytré karantény" v souvislosti

[5] Firma Amazon podle všeho nositelné zařízení rozpoznávající emoce (byť za pomoci analýzy tónu hlasu) dokonce skutečně vyvíjí [Day 2019].

s pandemií COVID-19. Chytrá karanténa (v případě souhlasu nakaženého) využívá data mobilních operátorů a data z platebních karet bankovních společností k vytvoření „vzpomínkové mapy" míst, na kterých se nakažený pohyboval. Do stejného systému zapojená mobilní aplikace eRouška zase pomocí Bluetooth zaznamenává ostatní uživatele aplikace v okolí a ulehčuje tak identifikaci těch, kdo byli s nakaženým v epidemiologicky významném kontaktu [MZČR 2020]. Totalitní režimy mají pochopitelně nespornou výhodu v tom, že disponují v podstatě neomezeným přístupem k datům svých občanů, jejichž obavami o soukromí a souhlasem se nemusí zdaleka tolik zatěžovat. Naproti tomu nasazení podobných opatření (na jiné než dobrovolné bázi) bývá ve zbytku světa terčem kritiky – např. provalení sběru dat NSA z úst whistleblowera Edwarda Snowdena vedlo k domácímu i mezinárodnímu skandálu a legislativním změnám [MacAskill 2015].

Digitální stopa z pohledu výzkumníka

Soubor všech dat a metadat, která v digitálních systémech zůstávají jako otisk aktivity jejich uživatelů, nazýváme pro potřeby tohoto článku „digitální stopa". V existující akademické literatuře se o stejných datech ovšem dočteme pod různými jmény. Někteří autoři hovoří o „digital trace data" [Jungherr 2015], jiní o „digital footprints" [Golder, Macy 2014] nebo „social transactional data" [Savage, Burrows 2007]. Další zůstávají u zastřešujícího termínu „big data" a až přídavným „in social science" [Foster et al. 2016] specifikují, že se nejedná například o data z částicových urychlovačů. Olshannikova et al. [2017] pro velká data ze sociálních sítí a různých forem lidské digitální komunikace používají termín „big social data".[6] Zachovat v názvu co největší propojení s výzkumnou tradicí big data může být užitečné i z toho důvodu, že část problémů a metod jejich řešení je s technickými obory buď zcela společná, nebo alespoň stojí na společném základu. Dekódování sémantiky lidmi vytvořeného obsahu[7] (zpracování přirozeného jazyka, rozpoznání řeči atd.) je problém do značné míry specifický pro lidmi vytvořený obsah, ale digitální stopa zahrnuje i množství dat vytvořených strojově,[8] kde bude průnik metod s technickými obory opravdu velký. Kromě různých názvů pro novou kategorii výzkumných dat vzniká i množství metodologických a myšlenkových směrů, které zdůrazňují proměny, jež s sebou nová data přináší.

[6] A dělí je na data (1) digitální sebereprezentace (profilová data a publikovaný obsah), (2) technologicky zprostředkované komunikace (soukromé zprávy i veřejné příspěvky v diskuzích), (3) digitálních vztahů (přátelství, odběry, sledování…) [Olshannikova et al. 2017: 14].

[7] Lidé v prostředí platforem tzv. Webu 2.0 představují nejen spotřebitele, ale i primární tvůrce obsahu (videa na YouTube, ale i komentáře pod nimi, fotografie na Instagramu, tweety na Twitteru atd.).

[8] Např. metadata typu kdo, kdy, kde, s kým atd., označovaná také jako „transakční data" [Beer, Burrows 2013].

„Computational social science" [Lazer et al. 2009] klade důraz na nezbytnost propojení metod informatiky a sociálních věd pro efektivní práci v oblasti big social data. „Forensic social science", kterou Goldberg chápe jako „pečlivý sběr důkazů z nestrukturované digitální stopy jako prostředek k vytváření nových teorií", podle něj dokonce nabízí pro sociologii příležitost konečně se osvobodit „z okovů testování hypotéz" [Goldberg 2015: 1].

Základní definice big data hovoří o datech, která standardní software a hardware není schopen v rozumném čase zpracovat [Manovich 2011] pro jejich absolutní velikost. Z pohledu výzkumných možností jsou ale big data zajímavá i velikostí relativní: namísto reprezentativních vzorků je totiž u big data teoreticky možné pracovat s daty celé příslušné populace (např. telefonní operátor má pro analýzy k dispozici metadata všech hovorů a zpráv, které se v jeho síti odehrály; banka údaje o všech transakcích svých klientů; Facebook veškeré interakce svých uživatelů). Kompletního zastoupení populace dosáhneme zatím[9] samozřejmě spíše u populací takto specificky vymezených (zákazníci telefonního operátora, klienti banky, uživatelé Facebooku). S rostoucí monopolizací trhu a vyhlídkou na téměř univerzální využívanost některých digitálních služeb však není do budoucna nepředstavitelný ani scénář dostupnosti dat o takřka libovolné kompletní populaci.

Značná část sociologických výzkumů stojí na datech získaných ze subjektivních výpovědí zkoumaných osob. Ty však mohou být zdrojem četných zkreslení. Digitální stopa naproti tomu poskytuje také strojové záznamy o aktivitách pořízené v momentě jejich konání. Namísto mlhavého vzpomínání, kde subjekt v daný den a danou hodinu byl, nám jako výzkumníkům může zpřístupnit záznamy své GPS polohy. Namísto odhadování, s kolika přáteli se uplynulý rok bavil o problematice fake news, nás může nechat strojově analyzovat svou historii zpráv. Namísto snahy artikulovat na místě názor na celou politickou scénu, nám může k analýze poskytnout své jednotlivé interakce s příspěvky politiků na sociálních sítích a jejich postupný vývoj za posledních 10 let.

Nedocenitelnou vlastností digitální stopy je i to, že umožňuje vyhnout se zkreslením vyvolaným tradičními výzkumnými stimuly, neboť je tvořena daty vzniklými v prostředí přirozených sociálních interakcí, a nikoli v rámci laboratorního experimentu. Data digitální stopy tak nepředstavují reakce na otázky výzkumníka, ale otisky chování člověka v jeho dnes již víceméně přirozeném prostředí. Zajímá-li se výzkumník o experimentální, a nikoli observační data, nabízí se možnost konstruovat v online prostředí v podstatě terénní experiment[10] a následně využít schopnosti platformy poskytnout bohatá doprovodná data v podobě digitální stopy. Například Bond et al. [2012] takto do experimentu s šířením

[9] V případě internetové služby využívané zpočátku primárně mladými a technologicky zdatnými lidmi zůstane zbytek populace skryt za „digitální propastí", která následnou zobecnitelnost výsledků může omezit.

[10] Konkrétní podobu „terénního výzkumu online" navrhují např. Parigi, Santana a Cook [2017].

volebního chování na Facebooku zapojili jinak těžko představitelných 61 milionů subjektů. Archivy digitální stopy mohou také být nebývale bohatým zdrojem tzv. „přirozených experimentů" a čím více se realita provozovatelů digitálních platforem typu Facebook bude blížit stavu „sbíráme všechna data o všech", tím spíš bude v archivech možné naleznout přirozeně se vyskytující výsledky různých experimentálních scénářů s přesně definovanými podmínkami.

Digitální stopa má krom silných stránek ale samozřejmě i mnohé slabé a ve srovnání s tradičními metodami hůře prozkoumané. Z pohledu výzkumníka znamená v jistém smyslu analýzu sekundárních dat, neboť jejich primární sběr mají vždy v rukou nástroje platformy. „Sběr dat" totiž platforma provádí v momentě samotné interakce a data jsou následně vyvolávána z databáze, kdykoli je třeba je zobrazit. Pokud tedy například komentář pod statusem na sociální síti zobrazujeme s odstupem několika let (nebo třeba jen na jiném zařízení), můžeme jej vidět ve zcela jiném prostředí, než v jakém jej autor psal a můžeme mylně předpokládat jiné možnosti interakce, než jaké měl autor k dispozici.[11] Výzkumník se tak musí vzdát značné části kontroly nad sběrem dat, a tedy i desítky let studovaných, standardizovaných metod empirické sociologie, které návrh tradičního výzkumu formují od samého počátku. Expertiza zde už nespočívá ve schopnosti správně položit respondentovi otázku, ale spíše porozumět tomu, jak neustále se měnící prostředí a nástroje konkrétního digitálního média formují samotné chování člověka.

I od ideálu nezkreslených dat se digitální stopa může velice vzdálit. Například výzkumy pracující s údaji subjektů získanými z profilů na sociálních sítích obvykle v základních demografických ukazatelích (které nezřídka rozhodují už o zapojení do vzorku) spoléhají na informace, které subjekt v profilu sám uvedl. Tím se nejen vracejí ke zkreslením subjektivních výpovědí, ale dokonce potenciálně přidávají další, protože ve srovnání s neveřejnými dotazníky klasického výzkumu mohou veřejné profilové údaje být zkreslené právě už samotným umístěním ve „výkladní skříni" [Hogan 2010]. Například studie, která se snaží sledovat studenty kalifornské univerzity, tak ve skutečnosti sleduje spíše uživatele Facebooku, kteří se na svém profilu k danému univerzitnímu studiu hlásí [Golder, Macy 2014: 1064].[12] Podobně se může zkomplikovat i představa o vyhnutí se zkreslením vyvolaným výzkumnými stimuly: Pokud si je uživatel vědom, že platforma veškerou jeho aktivitu zaznamenává, může její systémy už automaticky obcházet a např. citlivý obsah prohlížet v anonymních oknech a speciálních prohlížečích, citlivá témata rozebírat v koncově šifrovaných konverzacích atd.

Bylo by naivní tvářit se, že sociální sítě a další digitální technologie jsou transparentní a chování svých uživatelů neovlivňují. I kdybychom však z tohoto důvodu chtěli digitální stopu odmítnout jako cestu k tradičnímu, „offline" socio-

[11] Například tlačítko „to se mi líbí" přibylo k facebookovému obsahu v roce 2009 a v roce 2016 jej nahradilo celé spektrum možných „reakcí".

[12] Problematice validity a reliability Big Data se hlouběji věnují například zmínění Golder a Macy [2014] nebo Liu et al. [2016].

logickému poznání, neutečeme před skutečností, že se do online prostředí přesouvá stále větší část běžného života miliard lidí, a ty nejtriviálnější změny v jeho fungování tím pádem mohou mít na lidské životy větší dopad než mnohá roky debatovaná politická rozhodnutí. Tyto dopady je třeba studovat a využití nových metod se zde vyhnout nelze. Už Mills varuje před „fetišizací metody", mající za následek, že výběr zkoumaných témat je svým způsobem předem formován pevně daným metodologickým repertoárem [Mills (1959) 2000: 121–122; Gane 2011: 159], a ať už tradiční metody tvoří jakkoli silný rámec, zůstane-li akademický výzkumník odkázaný na nákladná n = 500 dotazníková šetření a časově náročné n = 20 hloubkové rozhovory, zatímco výzkumník komerční bude mít během několika kliknutí k dispozici strojově zpracovatelná data spočívající v kompletní historii chování milionových populací na úrovni jednotlivých interakcí, může mezi oběma světy vzniknout nepřekonatelná propast.

(Ne)dostupnost digitální stopy pro výzkum

Nejelegantnější cestou k datům pro výzkumníka v online prostředí je bezpochyby přímá spolupráce s platformou, jejíž data chce analyzovat. V případě velkých nadnárodních firem je však tato možnost dostupná výhradně projektům, které probíhají pod jejich střechou, v rámci jejich výzkumné agendy, nebo na kterých se alespoň přímo podílejí jejich zaměstnanci. Cesta, jak z pozice nezávislého výzkumníka požádat například Facebook o poskytnutí výzkumných dat na vlastní projekt a uspět, v současnosti prakticky neexistuje. Tato situace výzkumníky nevyhnutelně dělí na ty s daty a ty bez dat, přičemž druhá skupina nejenže nemá možnost výzkum té první reprodukovat, ale mnohdy ani kriticky zhodnotit jeho metodologii [Boyd, Crawford 2012: 673]. V momentě, kdy je přístup k datům privilegiem, které může poskytovatel kdykoli odebrat, výzkumník navíc nemůže riskovat kladení takových otázek, které by pro poskytovatele dat mohly být nepříjemné [ibid.: 674].

Naopak nejméně elegantní, zato nejobecněji použitelnou metodu sběru na webu dostupných dat představuje tzv. „scraping" neboli procházení obsahu webových stránek a jeho ukládání pro pozdější analýzu.[13] Takový postup má pro obě strany pochopitelně některé nevýhody: pro výzkumníka je to především potřeba specializovaných programátorských dovedností nutných pro vytvoření nástrojů, které budou specificky formátovaná data pomáhat sbírat a zpracovávat; nižší rychlost a menší spolehlivost načítání obsahu skrz web a v neposlední řadě i nutnost být si vědom podmínek užívání služby (které mohou např. automatizovaný sběr výslovně zakazovat). Pro provozovatele platformy je to pak třeba větší zátěž serverů nebo nemožnost ovlivnit, který obsah bude dostupný strojově a který pouze skutečnému uživateli.

[13] Ať už se jedná o procházení ruční, nebo automatizované a ať už se využívá tradičních webových prohlížečů, nebo přímé komunikace přes http protokol.

Standardem dnešní doby ve výzkumném sběru dat z velkých webových platforem je využití tzv. API,[14] skrz které provozovatel platformy zpřístupní ta data, která zpřístupnit chce,[15] v množství, které je zpřístupnit ochoten, a zároveň ve formátu a rychlostí, které jsou velkou výhodou i pro výzkumníka. V případě platforem s individuálními profily pak může o zpřístupnění vlastních údajů (typicky pro potřeby aplikace, která je chce využít) mnohdy rozhodnout v daném momentě přímo uživatel sám.

Jedním z významných příkladů výzkumného sběru digitální stopy ze sociálních sítí s pomocí API (a s explicitním souhlasem každého participanta) byla facebooková aplikace „myPersonality" Davida Stillwella, kterou se následně inspirovala i zmíněná kontroverzní Cambridge Analytica. Mezi roky 2007–2012 umožňovala uživatelům vyplnit osobnostní test, který následně zdarma vyhodnotila. Součástí testu byl i souhlas se sdílením digitální stopy a demografických údajů z Facebooku, který udělilo cca 40 % uživatelů aplikace z celkových šesti milionů. Výsledný dataset v sobě tedy kombinoval demografická data subjektů, odpovědi v testu osobnosti a množství digitálních stop v čele se seznamem oblíbených facebookových stránek uživatele. Tato data byla následně zpřístupňována dalším výzkumníkům až do roku 2018, kdy jejich poskytování autoři v návaznosti na tlak vyvolaný aférou Cambridge Analytica ukončili. Článek autorů z roku 2015 uvádí, že dataset využilo minimálně 200 výzkumníků ze 100 různých akademických institucí [Kosinski et al. 2015] a aktuální stránka projektu odkazuje na 57 vybraných publikací [myPersonality n.d.].

Jiná vysokoprofilová výzkumná aplikace Netvizz [Rieder 2013] uživatelská data nesbírala, ale naopak k nim uživateli skrz API Facebooku umožňovala přistupovat a analyzovat tak například veřejně dostupné příspěvky médií nebo politiků. Dlouhou dobu patřila k nejoblíbenějším analytickým nástrojům pro facebooková data a v době publikace tohoto textu nasbírala prostřednictvím příslušného akademického článku přes 500 citací.[16]

Na skandály, jako bylo komerční zneužití facebookových dat firmou Cambridge Analytica, a na regulační snahy v podobě evropského GDPR však sociální sítě v čele právě s Facebookem odpovídají skokovým omezováním dostupnosti

[14] „Application Programming Interface" neboli „rozhraní pro programování aplikací" je standardizované rozhraní dané platformy (například Facebooku, YouTube, Twitteru) určené primárně pro kontrolovaný přístup k jejím funkcím a datům z aplikací vývojářů třetích stran.

[15] Výzkumníci pracující s API nebo scrapingem jsou vždy omezeni na data z kategorie veřejně přístupných. Pouze výzkumy prováděné v přímé spolupráci s platformou mohou mít k dispozici celou digitální stopu uživatele tak, jak ji platforma pro interní potřeby sbírá. Například výzkum autorů Chenga, Burke a Goetz Davis [2019] využívá neveřejné metriky Facebooku jako počet návštěv stránky, čas strávený na stránce, frekvence posílání zpráv, interakce s blízkými přáteli, množství zobrazených notifikací, zobrazené příspěvky a jejich obsah atd.

[16] Podle výsledků vyhledávače Google Scholar ze dne 29. 4. 2020.

dat z API, až se tato mladá a perspektivní větev výzkumu postupně dostává do stavu blízkého nepoužitelnosti. Jelikož se obchodní model internetových gigantů točí právě kolem analýzy uživatelských dat, jejich znepřístupnění ve jménu vyhovění regulátorům a tlakům veřejnosti pro ně představuje spíše obchodní výhodu, zatímco pro nezávislý výzkum znamená v lepším případě návrat ke scrapingu (který výzkum při troše neopatrnosti snadno dostane za hranici tolerance etických komisí), v horším neproveditelnost.

Část akademické obce pochopitelně volá po přehodnocení takové strategie [Bruns 2018; Rieder 2018; Brooks 2018]. Za jistou formu odpovědi na toto volání pak můžeme považovat např. spolupráci Facebooku s organizací Social Science One. Společná iniciativa „Social Media and Democracy Research Grants" [Schrage, Ginsberg 2018] pro vybrané akademiky představovala první možnost systematického přístupu k anonymizovaným datům z Facebooku. Spolupráce má však zatím pevně danou výzkumnou agendu, která vychází ze závazků Facebooku vůči Kongresu Spojených států amerických (vliv sociálních médií na volby) a přístup k datům je podmíněn úspěchem v komplikovaném grantovém řízení. Od ohlášení iniciativy po vyhlášení prvních úspěšných žadatelů navíc uplynul rok, takže v tuto chvíli nezbývá než doufat, že se jedná spíše o pilotní projekt, jehož cílem je otestovat nový, poměrně komplikovaný způsob zpřístupňování dat,[17] neboť hrstku projektů ročně (navíc se stejným výzkumným zaměřením) lze jen těžko považovat za adekvátní akademický přístup k jednomu z největších potenciálních zdrojů dat pro sociální vědy.

Příklady výzkumného využití digitální stopy

Už Golder a Macy ve své přehledové studii z roku 2014 uvádějí, že pole výzkumu digitální stopy je příliš rychle rostoucí a rozsáhlé na to, aby bylo možné vytvořit jeho celkový přehled. Dnes to samozřejmě platí dvojnásob. V tomto oddílu přinášíme krátká shrnutí několika studií, které dobře ilustrují potenciál doposud zmíněných využití digitální stopy na datech z jedné ze dvou nejčastěji zkoumaných sociálních sítí (Facebook, Twitter). Tam, kde je to možné, připojujeme ke zmíněným tématům odkazy na relevantní výzkumy z českého prostředí. Ne vždy se jedná o studie výhradně sociologické, neboť digitální stopa má často silnější výzkumnou tradici v příbuzných společenských vědách (např. v behaviorální a sociální psychologii), kterou by byla škoda ignorovat mimo jiné i proto, že stejná data mohou sloužit i ke studiu otázek sociologických.

[17] Navrhovaný nový model spolupráce mezi akademickou a komerční sférou v článku rozebírají King a Persily [2019], kteří stojí za organizací Social Science One, jež přístup k facebookovým datům zprostředkovává. Zpřístupňovaný dataset obsahuje speciálně kalibrovaný náhodný šum, který byl přidán jako matematická záruka zachování soukromí participantů a vyžaduje tedy speciální metody zpracování, aby se zabránilo chybám měření vedoucím ke statistickým zkreslením [Evans, King 2020].

Jako první se zaměříme na studie postavené na datasetu myPersonality, jenž se stal inspirací pro kontroverzní aktivity poradenské společnosti Cambridge Analytica, které potom spolu s aktivitami americké NSA a evropskou legislativou GDPR nejvíce formovaly všeobecné povědomí o digitální stopě a ovlivnily její výzkumnou dostupnost. Kosinski, Stillwell a Graepel [2013] na datech z myPersonality ukázali, že i základní veřejně dostupné informace z facebookových profilů mohou být použity k automatizované a poměrně přesné predikci potenciálně citlivých osobních údajů (např. pohlaví, věk, etnicita, sexuální orientace, náboženské vyznání, politické postoje, inteligence, užívání návykových látek, neúplná rodina atd.). Strojové učení zde pracovalo s výsekem 58 000 subjektů zmíněného datasetu. Výsledný prediktivní model potom ze samotného seznamu oblíbených stránek dosáhl například 88% úspěšnosti rozpoznání homosexuality, 95% úspěšnosti rozpoznání bělocha od Afroameričana nebo 85% úspěšnosti rozlišení voliče demokratů/republikánů.

Navazující studie [Youyou, Kosinski, Stillwell 2015] srovnávala přesnost posuzování osobnostních rysů (tzv. „Big Five" – pětifaktorového modelu osobnosti[18]) mezi člověkem a počítačovým prediktivním modelem, který měl k dispozici pouze digitální stopu zkoumaného jedince. Autoři ukázali nejen, že je z digitální stopy hodnotit osobnost člověka možné, ale dokonce na základě naměřených výsledků tvrdí, že hodnocení jejich prediktivního modelu je přesnější než hodnocení partnerů, přátel a kolegů subjektu. U některých hodnot prediktivní model dosahoval dokonce větší shody s výsledkem osobnostního testu, než jakého dosahovalo sebehodnocení subjektu samotného.

Matz et al. [2017] se rozhodli ukázat sílu přizpůsobování reklamní komunikace na základě psychologického profilu adresáta vytvořeného z jeho veřejně dostupné digitální stopy a opět využili data z aplikace myPersonality. V několika samostatných experimentech testovali mimo jiné reklamní sdělení přizpůsobené pro extroverta/introverta (a dosáhli 50% nárůstu prodejů oproti kontrolní skupině) nebo reklamu na mobilní aplikaci přizpůsobenou podle jednoho z faktorů Big Five – uživatelovy otevřenosti k novým zkušenostem (a dosáhli až 40% nárůstu prokliků). Oba experimenty pracovaly s cílením na základě jediné oblíbené stránky subjektu (namísto jejich kompletního seznamu), takže se jedná o výsledky co do možností dané metody velice konzervativní. V závěrečné diskusi autoři poukazují jak na možnosti pozitivního využití psychologického cílení komuni-

[18] Pětifaktorový model osobnosti má v sociálních vědách nezanedbatelnou tradici [např. Selfhout et al. 2010; Gerber et al. 2011; v prostředí české sociologie např. Lyons 2017; Anýžová 2018]. Osobnostní rysy se, jak píše Anýžová, „zpravidla neměří v běžných reprezentativních sociologických výzkumech z důvodu příliš rozsáhlé standardizované baterie výroků (v psychologii inventáře), která neúnosně zatěžuje již tak často předimenzované dotazníky" [2018: 668]. Obohacení dat osobnostním profilem zkoumaného jedince tam, kde vyplňování dlouhých baterií otázek není prakticky proveditelné, může zajistit právě analýza digitální stopy a otevřít tak další výzkumné možnosti.

kace (nasměrování subjektů jevících známky deprese na odbornou pomoc nebo literaturu), tak na možnost zneužití slabin subjektu (např. pomocí reklamy na kasino cílené na jedince jevící rysy spojené s patologickým gamblerstvím) a zmiňují i v té době ještě nepotvrzená podezření o takovéto manipulaci kampaně amerických prezidentských voleb v roce 2016, tedy cílenou snahu o „demobilizaci" voličů Hillary Clintonové.

V předchozích oddílech jsme zmínili možnost konstrukce experimentů v online prostředí. Významným příkladem takového postupu je studie provedená v přímé spolupráci Facebooku a Kalifornské univerzity [Bond et al. 2012], která se věnovala šíření volebního chování v prostoru sociální sítě. Autoři ukázali jednak efekt společenského přenosu volebního chování, jednak i rozdílný vliv silných a slabých sociálních vazeb na něj. Uvádějí, že ač tradiční experimenty s mobilizací voličů vykazují spíše sotva měřitelný dopad na skutečnou volební účast, online sociální sítě – vzhledem k potenciálu zasáhnout nebývale velké množství lidí – mohou přesto vyvolat milionové rozdíly v účasti a výsledky významně ovlivnit. V předkládaném experimentu výzkumníci uvádějí, že v populaci 61 milionů zacílených subjektů docílili 340 tisíc volebních hlasů v urnách navíc. Politické participaci na sociálních sítích, byť formou „tradiční" analýzy digitální stopy spíše než experimentu, se věnují i české výzkumné projekty. Například na genderové nerovnosti v politické participaci na Facebooku se s pomocí obsahové analýzy komentářů a příspěvků zaměřili Vochocová, Mazák a Štětka [2016].

Zmiňovali jsme, že i kdybychom chtěli digitální stopu odmítnout jako cestu k „offline" sociologickému poznání, nevyhneme se jí při studiu jevů, které přicházejí s novými způsoby komunikace. Dobrým příkladem může být otázka tzv. názorových bublin („filter bubbles",[19] případně „echo chambers"), tedy stavu, kdy se člověk uzavírá v názorově homogenním prostředí, které může vést k jeho dalšímu vyhraňování. Například Zelenka et al. [2016] v projektu „Češi v zajetí sociálních bublin" na digitální stopě uživatelů Facebooku ukazovali, jak zásadně se liší zdroje informací, které sledují příznivci tradičních a antisystémových stran. Stejně tak Šlerka a Šisler [2018] v debatě o uprchlické krizi na českém Facebooku identifikovali přinejmenším čtyři distinktivní názorová seskupení, která se u stejného mediálního obsahu potkávala zcela minimálně. Obě zmíněné studie využívaly digitální stopu v podobě uživatelských interakcí s veřejným obsahem facebookových stránek.

Velmi silná výzkumná tradice existuje kolem mikroblogovací platformy a sociální sítě Twitter, neboť ta svůj obsah (na rozdíl od relativně uzavřeného Facebooku) v základním nastavení zpřístupňuje veřejně a i podmínky výzkum-

[19] Termín „filter bubble" zdůrazňuje v problematice především roli algoritmů. Ty v prostředí sociálních médií fungují skutečně jako „filtr" a rozhodují o tom, jaký obsah uživateli ukázat. Typicky tak činí právě na základě digitálních stop a vypočítané pravděpodobnosti, co by se uživateli mohlo líbit, případně s jakým obsahem bude nejpravděpodobněji interagovat.

ného přístupu k datům má podstatně otevřenější. Na rozdíl od Facebooku zde výzkumník stále může využívat API, a ačkoli za plnohodnotný přístup k datům musí platit, i zdarma poskytovaný přístup může být dostatečný pro některé menší studie a pro většinu studií pilotních. Významné uplatnění nachází Twitter tradičně v oblasti studia politické komunikace, čemuž se detailně věnuje například publikace *Analyzing Political Communication with Digital Trace Data* [Jungherr 2015]. Bright [2018] zase na interakčních datech z Twitteru ukazuje, že vzájemná komunikace mezi příslušníky politického středu je podstatně častější než mezi příslušníky opačných frakcí na extrémních koncích politického spektra. Pro výzkumníky zabývající se dezinformacemi a propagandou na sociálních sítích mohou být zajímavé také volně dostupné datasety státem financovaných informačních operací, které Twitter odhaluje a pravidelně zpřístupňuje v rámci „Transparency report" [Twitter, Inc. 2020]. Na datech z Twitteru je pochopitelně možné studovat i jevy zcela apolitické, jako například vývoj pocitu štěstí a spokojenosti ve společnosti, který s pomocí 4,6 miliard tweetů a analýzy sentimentu mapovali Dodds et al. [2011]. Výsledky jejich tříleté analýzy názorně ukazují zmiňovanou granularitu digitální stopy, která výzkumníkům umožnila sledovat vývoj zkoumaného jevu zároveň na úrovni let, měsíců v roce, dní v týdnu, ale i hodin, stejně jako identifikovat v datasetu celospolečenské momenty pozitivně i negativně významných událostí.

Závěr

Cílem tohoto textu bylo představit čtenáři koncept digitální stopy, tedy jakéhosi pomyslného souboru všech dat a metadat, která v digitálních systémech zůstávají jako otisk každodenních aktivit uživatelů. Z komerčního pohledu tvoří sběr a analýza digitální stopy základ obchodního modelu současného internetu, kde nejčastěji slouží k co nejpřesnějšímu vyprofilování uživatele a následně personalizovanému filtrování obsahu a cílení reklamy. Z pohledu složek státního aparátu je digitální stopa užitečným nástrojem dohledu, kde profil jedince už dnes pomůže rozhodnout o vpuštění do země, zabrání mu v nákupu letenek a jízdenek, případně poslouží jako důkazní materiál v trestněprávním řízení. Z pohledu akademického výzkumu představuje digitální stopa zdroj dat o velkých populacích, vyznačující se zároveň bezprecedentní granularitou: detailní záznamy o pohybu, společenských interakcích, kulturní spotřebě nebo třeba zdravotních ukazatelích miliard jednotlivců.

Rozdíl mezi objemem dat dostupným provozovatelům služeb a nezávislým výzkumníkům s postupem času propastně narůstá. Provozovatelé jich sbírají stále větší množství a stále méně jich zpřístupňují, k čemuž jim paradoxně pomáhají i aféry a regulace. Za situace, kdy v důsledku afér typu Cambridge Analytica stojí akademický přístup k datům a soukromí uživatelů zdánlivě v přímé opozici, zvítězí nevyhnutelně soukromí – tím spíš, když provozovateli umožní upevnit datový monopol, který tvoří páteř jeho obchodního modelu. Velké praktické riziko

pro nezávislý výzkum digitální stopy představují i samotné změny v dostupnosti dat, které jsou často nepředvídatelné a neumožňují plánovat výzkum, nemá-li výzkumník v rukou již hotový dataset. Pro představu – metodologie většiny výše uvedených nezávislých studií Facebooku bude dnes, několik let po jejich provedení, nereplikovatelná, protože stejná data již není možné sebrat. Facebook na sílící kritiku svého postupu vůči výzkumníkům sice reaguje, ale zatím není jasné, jestli jde o skutečnou snahu spolehlivě anonymizovaná data dlouhodobě a v dostatečné míře zpřístupnit, nebo pouze o vyhovění Kongresu Spojených států amerických a symbolický PR krok v podobě umožnění několika málo studií jedné kritizované oblasti svého působení. Facebook pochopitelně není jediným zdrojem dat, byť trend jejich uzavírání je v různé míře patrný takřka všude. Hrozí tak, že časy dostupnosti nezměrného množství dat (jak tomu bylo v době sběru datasetů myPersonality a Cambridge Analytica) skončí dřív, než velká část výzkumníků v oblasti sociálních věd zaregistruje, že začaly.

Digitální stopa představuje jeden z potenciálně nejvýznamnějších typů dat pro výzkum v sociálních vědách a online sociální sítě jeden z jeho nejvýznamnějších zdrojů. S oběma se ale pochopitelně pojí i četná metodologická úskalí. Někdy jde o staré problémy v novém kabátě, k jejichž řešení je málokdo povolanější než právě odborníci na empirický výzkum z řad sociologů, jindy o problémy zcela nové, které vyžadují jak silnou mezioborovou spolupráci, tak spolupráci s komerční sférou. Je důležité, aby studium digitální stopy mělo v akademickém prostředí oporu i na vyšších místech, než jaká zaujímají technologicky schopní studenti a doktorandi, z nichž navíc většina po dokončení studia zamíří spíše dále posílit komerční sektor. Jedině silná akademická základna má šanci řešit problém pro studium digitální stopy zcela existenční a dostatečně hlasitě se zasadit o rozumné, oficiální zpřístupnění těchto dat pro výzkumníky a akademické instituce, které by umožnilo digitální stopu dlouhodobě a s minimem vzájemných rizik studovat. Do té doby výzkumníkům nezbývá než využít cest, které ještě nebyly uzavřeny, platforem, které data ještě zpřístupňují, případně se vrátit k primitivnějším a časově náročnějším metodám, než je pohodlný a nekontroverzní sběr skrz API. Dveře největšího repozitáře sociologických dat v historii již nejsou tak docela dokořán, ale stále zůstávají pootevřené.

Jakub Sedláček *učí na Studiích nových médií Filozofické fakulty Univerzity Karlovy a zároveň je doktorandem Institutu sociologických studií na Fakultě sociálních věd. Na datech ze sociálních sítí zkoumá společenskou roli celebrity a vliv influencerů na mladé lidi.*

Literatura

Anýžová, P. 2018. „Vztah mezi individuálními charakteristikami a vysokoškolským vzděláním: role osobnosti, fyzické atraktivity a sebehodnocení.“ *Sociologický časopis / Czech Sociological Review* 54 (5): 667–697, https://doi.org/10.13060/00380288.2018.54.5.419.

Beer, D., R. Burrows. 2013. „Popular Culture, Digital Archives and the New Social Life of Data.“ *Theory, Culture & Society* 30 (4): 47–71, https://doi.org/10.1177/0263276413476542.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, J. H. Fowler. 2012. „A 61-million-person Experiment in Social Influence and Political Mobilization.“ *Nature* 489 (7415): 295–298, https://doi.org/10.1038/nature11421.

Boyd, D., K. Crawford. 2012. „Critical Questions for Big Data: Provocations or a Cultural, Technological, and Scholarly Phenomenon.“ *Information, Communication & Society* 15 (5): 662–679, https://doi.org/10.1080/1369118X.2012.678878.

Bright, J. 2018. „Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism.“ *Journal of Computer-Mediated Communication* 23 (1): 17–33, https://doi.org/10.1093/jcmc/zmx002.

Brooks, C. 2018. „It's Time for Facebook to Share More Data with Researchers.“ *Wired* [online]. [cit. 11. 5. 2019]. Dostupné z: https://www.wired.com/story/its-time-for-facebook-to-share-more-data-with-researchers/.

Bruns, A. 2018. „Facebook Shuts the Gate after the Horse Has Bolted, and Hurts Real Research in the Process.“ *Internet Policy Review* [online]. [cit. 3. 4. 2019]. Dostupné z: https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786.

Cadwalladr, C., E. Graham-Harrison. 2018. „Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach.“ *The Guardian* [online]. [cit. 3. 5. 2019]. Dostupné z: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

Constine, J. 2018. „A Flaw-by-flaw Guide to Facebook's New GDPR Privacy Changes.“ *Techcrunch* [online]. [cit. 20. 5. 2019]. Dostupné z: https://techcrunch.com/2018/04/17/facebook-gdpr-changes/.

Crandall, D. J., L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, J. Kleinberg. 2010. „Inferring Social Ties from Geographic Coincidences.“ *Proceedings of the National Academy of Sciences of the United States of America* 107 (52): 22436–22441, https://doi.org/10.1073/pnas.1006155107.

Day, M. 2019. „Amazon Is Working on a Device That Can Read Human Emotions.“ *Bloomberg* [online]. [cit. 28. 5. 2019]. Dostupné z: https://www.bloomberg.com/news/articles/2019-05-23/amazon-is-working-on-a-wearable-device-that-reads-human-emotions.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, C. M. Danforth. 2011. „Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter.“ *PLoS ONE* 6 (12), https://doi.org/10.1371/journal.pone.0026752.

Evans, G., G. King. 2020. „Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset.“ Working Paper. *Gary King* [online]. Dostupné z: https://j.mp/38NrmRW.

Foster, I., R. Ghani, R. S. Jarmin, F. Kreuter, J. Lane. 2016. „*Big Data and Social Science: A Practical Guide to Methods and Tools*.“ Chapman and Hall/CRC, https://doi.org/10.1201/9781315368238.

Gane, N. 2011. „Measure, Value and the Current Crises of Sociology.“ *Sociological Review* 59 (suppl. 2): 151–173, https://doi.org/10.1111/j.1467-954X.2012.02054.x.

Gerber, A. S., G. A. Huber, D. Doherty, C. M. Dowling. 2011. „The Big Five Personality Traits in the Political Arena." *Annual Review of Political Science* 14 (1): 265–287, https://doi.org/10.1146/annurev-polisci-051010-111659.
Glazer, E., D. Seetharaman, A. Andriotis. 2018. „Facebook to Banks: Give Us Your Data, We'll Give You Our Users." *The Wall Street Journal* [online]. [cit. 27. 4. 2019]. Dostupné z: https://www.wsj.com/articles/facebook-to-banks-give-us-your-data-well-give-you-our-users-1533564049.
Goldberg, A. 2015. „In Defense of Forensic Social Science." *Big Data and Society* 2 (2): 1–3, https://doi.org/10.1177/2053951715601145.
Golder, S. A., M. Macy. 2014. „Digital Footprints: Opportunities and Challenges for Online Social Research." *Annual Review of Sociology* 40: 129–152, https://doi.org/10.1146/annurev-soc-071913-043145.
Hill, K. 2018. „Facebook Is Giving Advertisers Access to Your Shadow Contact Information." *Gizmodo* [online]. [cit. 20. 4. 2019]. Dostupné z: https://gizmodo.com/facebook-is-giving-advertisers-access-to-your-shadow-co-1828476051.
Hogan, B. 2010. „The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online." *Bulletin of Science, Technology & Society* 30 (6): 377–386, https://doi.org/10.1177/0270467610385893.
Cheng, J., M. Burke, E. Goetz Davis. 2019. „Understanding Perceptions of Problematic Facebook Use: When People Experience Negative Life Impact and a Lack of Control." Pp. 1–13 in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, https://doi.org/10.1145/3290605.3300429.
Jungherr, A. 2015. *Analyzing Political Communication with Digital Trace Data*. Switzerland: Springer International Publishing, https://doi.org/10.1007/978-3-319-20319-5.
Kastrenakes, J. 2018. „Facebook Spoke with Hospitals about Matching Health Data to Anonymized Profiles." *The Verge* [online]. [cit. 7. 5. 2019]. Dostupné z: https://www.theverge.com/2018/4/5/17203262/facebook-medical-data-sharing-plan-healthcare.
King, G., N. Persily. 2019. „A New Model for Industry–Academic Partnerships." *Political Science & Politics* [online]. Cambridge University Press, 1–7, https://doi.org/10.1017/S1049096519001021.
Kosinski, M. 2018. „I Had Nothing to Do with Cambridge Analytica." *Michalkosinski.com* [online]. [cit. 17. 5. 2019]. Dostupné z: https://drive.google.com/file/d/1zRaTAx0mpRC0m7-3wQRaDPYTOGMdvNBt/edit.
Kosinski, M., S. C. Matz, S. D. Gosling, V. Popov, D. Stillwell. 2015. „Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines." *American Psychologist* 70 (6): 543–556, https://doi.org/10.1037/a0039210.
Kosinski, M., D. Stillwell, T. Graepel. 2013. „Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." *Proceedings of the National Academy of Sciences* 110 (15): 5802–5805, https://doi.org/10.1073/pnas.1218772110.
Kuo, L. 2019. „China Bans 23m from Buying Travel Tickets as Part of 'Social Credit' System." *The Guardian* [online]. [cit. 18. 5. 2019]. Dostupné z: https://www.theguardian.com/world/2019/mar/01/china-bans-23m-discredited-citizens-from-buying-travel-tickets-social-credit-system.
Landau, S. 2013. „Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations." *IEEE Security & Privacy* 11 (4): 54–63, https://doi.org/10.1109/MSP.2013.90.
Larson, Q. 2017. „I'll Never Bring My Phone on an International Flight Again. Neither Should You." *FreeCodeCamp* [online]. [cit. 24. 4. 2019]. Dostupné z:

https://www.freecodecamp.org/news/ill-never-bring-my-phone-on-an-international-flight-again-neither-should-you-e9289cde0e5f/.
Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne. 2009. „SOCIAL SCIENCE: Computational Social Science." *Science* 323 (5915): 721–723. https://doi.org/10.1126/science.1167742.
Liao, R. 2019. „A Government Propaganda App Is Going Viral in China." *Techcrunch* [online]. [cit. 15. 5. 2019]. Dostupné z: https://techcrunch.com/2019/02/01/china-propaganda-app/.
Liu, J., J. Li., W. Li., J. Wu. 2016. „Rethinking Big Data: A Review on the Data Quality and Usage Issues." *ISPRS Journal of Photogrammetry and Remote Sensing* 115: 134–142, https://doi.org/10.1016/j.isprsjprs.2015.11.006.
Lyons, P. 2017. „*Political Knowledge in the Czech Republic*." Praha: Sociologický ústav AV ČR, v. v. i.
Ma, A. 2018. „China Has Started Ranking Citizens with a Creepy 'Social Credit' System — Here's What You Can Do Wrong, and the Embarrassing, Demeaning Ways They Can Punish You." *Business Insider* [online]. [cit. 1. 5. 2019]. Dostupné z: https://www.businessinsider.c.om/china-social-credit-system-punishments-and-rewards-explained-2018-4.
MacAskill, E. 2015. „The NSA's Bulk Metadata Collection Authority Just Expired. What Now?" *The Guardian* [online]. [cit. 3. 5. 2019]. Dostupné z: https://www.theguardian.com/us-news/2015/nov/28/nsa-bulk-metadata-collection-expires-usa-freedom-act.
Manovich, L. 2011. „Trending: The Promises and the Challenges of Big Social Data." Pp. 460–475 in M. K. Gold. *Debates in the Digital Humanities*, https://doi.org/10.5749/minnesota/9780816677948.003.0047.
Matz, S. C., M. Kosinski, G. Nave, D. J. Stillwell. 2017. „Psychological Targeting as an Effective Approach to Digital Mass Persuasion." *Proceedings of the National Academy of Sciences* 114 (48): 12714–12719, https://doi.org/10.1073/pnas.1710966114.
Mills, C. W. (1959) 2000. *The Sociological Imagination*. New York: Oxford University Press.
myPersonality. n.d. „Publications." *MyPersonality.org* [online]. [cit. 6. 3. 2019]. Dostupné z: https://sites.google.com/michalkosinski.com/mypersonality/publications.
MZČR. 2020. „Chytrá karanténa." *Ministerstvo zdravotnictví ČR* [online]. [cit. 30. 4. 2020]. Dostupné z: https://koronavirus.mzcr.cz/chytra-karantena/.
Nowak, M., D. Eckles. 2016. *US20160283485A1: Determining User Personality Characteristics from Social Networking System Communications and Characteristics*. United States Patent and Trademark Office. Dostupné z: https://patents.google.com/patent/US20160283485A1/en.
Olshannikova, E., T. Olsson, J. Huhtamäki, H. Kärkkäinen. 2017. „Conceptualizing Big Social Data." *Journal of Big Data* 4 (1), https://doi.org/10.1186/s40537-017-0063-x.
Parigi, P., J. J. Santana, K. S. Cook. 2017. „Online Field Experiments: Studying Social Interactions in Context." *Social Psychology Quarterly* 80 (1): 1–19, https://doi.org/10.1177/0190272516680842.
Paul, K. 2019. „Libra: Facebook Launches Cryptocurrency in Bid to Shake up Global Finance." *The Guardian* [online]. [cit. 18. 6. 2019]. Dostupné z: https://www.theguardian.com/technology/2019/jun/18/libra-facebook-cryptocurrency-new-digital-money-transactions.
Penzenstadler, N., B. Heath, J. Guynn. 2018. „We Read Every One of the 3,517 Facebook Ads Bought by Russians. Here's What We Found." *USA Today* [online]. [cit. 20. 5. 2019]. Dostupné z: https://eu.usatoday.com/story/news/2018/05/11/what-we-found-facebook-ads-russians-accused-election-meddling/602319002/.

Quercia, D., M. Kosinski, D. Stillwell, J. Crowcroft. 2011. „Our Twitter Profiles, Our Selves: Predicting Personality With Twitter." Pp. 180–185 in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, https://doi.org/10.1109/PASSAT/SocialCom.2011.26.

Rieder, B., 2013. „Studying Facebook via Data Extraction." Pp. 346–355 in *Proceedings of the 5th Annual ACM Web Science Conference on – WebSci '13*, https://doi.org/10.1145/2464464.2464475.

Rieder, B. 2018. „Facebook's App Review and How Independent Research Just Got a lot Harder." *The Politics of Systems* [online]. [cit. 28. 3. 2019]. Dostupné z: http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/.

Rincon, J. A., A. Costa, P. Novais, V. Julian, C. Carrascosa. 2016. „*Using Non-invasive Wearables for Detecting Emotions with Intelligent Agents*." Pp. 77–84 in M. Graña, J. López-Guede, O. Etxaniz, Á. Herrero, H. Quintián, E. Corchado (eds.). *International Joint Conference SOCO'16-CISIS'16-ICEUTE'16. SOCO 2016, ICEUTE 2016, CISIS 2016. Advances in Intelligent Systems and Computing*, vol. 527. Springer, https://doi.org/10.1007/978-3-319-47364-2_8.

Savage, M., R. Burrows. 2007. „The Coming Crisis of Empirical Sociology." *Sociology* 41 (5): 885–899, https://doi.org/10.1177/0038038507080443.

Schrage, E., D. Ginsberg. 2018. „Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections." *Facebook Newsroom* [online]. [cit. 15. 5. 2019]. Dostupné z: https://newsroom.fb.com/news/2018/04/new-elections-initiative/.

Selfhout, M., W. Burk, S. Branje, J. Denissen, M. Van Aken, W. Meeus. 2010. „Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach." *Journal of Personality* 78 (2): 509–538, https://doi.org/10.1111/j.1467-6494.2010.00625.x.

Solon, O. 2017. „Facebook Can Track Your Browsing Even after You've Logged out, Judge Says." *The Guardian* [online]. [cit. 29. 3. 2019]. Dostupné z: https://www.theguardian.com/technology/2017/jul/03/facebook-track-browsing-history-california-lawsuit.

Statista. 2019. „Number of Daily Active Facebook Users Worldwide as of 1st Quarter 2019." *Statista* [online]. [cit. 24. 5. 2019]. Dostupné z: https://www.statista.com/statistics/346167/facebook-global-dau/.

Šlerka, J., V. Šisler. 2018. „Who Is Shaping Your Agenda? Social Network Analysis of Anti-Islam and Anti-immigration Movement Audiences on Czech Facebook." Pp. 61–85 in *Expressions of Radicalization*, https://doi.org/10.1007/978-3-319-65566-6_3.

Twitter, Inc. 2020. „Information operations." *Transparency Report* [online]. [cit. 1. 5. 2020]. Dostupné z: https://transparency.twitter.com/en/information-operations.html.

Vochocová, L., J. Mazák, V. Štětka. 2010. „Nic pro holky? Genderové nerovnosti v politické participaci na sociálních sítích." *Gender, rovné příležitosti, výzkum* 17 (2): 64–75, https://doi.org/10.13060/12130028.2016.17.2.283.

Youyou, W., M. Kosinski, D. Stillwell. 2015. „Computer-based Personality Judgments Are More Accurate than Those Made by Humans." *Proceedings of the National Academy of Sciences*, 112 (4): 1036–1040, https://doi.org/10.1073/pnas.1418680112.

Zelenka, J., T. Málek, M. Šulek, Š. Korčiš. 2016. „Češi v zajetí sociálních bublin." *Lidovky.cz* [online]. [cit. 8. 4. 2019]. Dostupné z: https://www.lidovky.cz/bubliny.aspx.