"API-based research" or how can digital sociology and journalism studies learn from the Cambridge Analytica affair

Tommaso Venturini, Richard Rogers

▶ **To cite this version:**

Tommaso Venturini, Richard Rogers. "API-based research" or how can digital sociology and journalism studies learn from the Cambridge Analytica affair. Digital Journalism , In press. hal-02003925

HAL Id: hal-02003925
https://hal.science/hal-02003925

Submitted on 1 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

"API-based research"
or how can digital sociology and journalism studies learn from the Cambridge Analytica affair

Tommaso Venturini & Richard Rogers

How to cite

Venturini, Tommaso, and Richard Rogers. 2019. "'API-Based Research' or How Can Digital Sociology and Digital Journalism Studies Learn from the Cambridge Analytica Affair." *Digital Journalism*, Forthcoming.

Keywords: digital sociology; computational social sciences; social media; Facebook ; digital ethnography

API-based research is an approach to computational social sciences and digital sociology based on the extraction of records from the datasets made available by online platforms through their application programming interfaces (or APIs). This type of research has allowed the collection of detailed information on large populations, thereby effectively challenging the distinction between qualitative and quantitative methods. Lately, access to this source of data has been significantly reduced after a series of scandals connected to the use of personal information collected through online platforms. The closure of APIs, however, can have positive effects, if it encourages researchers to reduce their dependence on mainstream platforms and explore new sources and ways to collect records on online interactions that are closer to the digital fieldwork.

Among other reasons, the 2016 US presidential election will be remembered for the revelation that its outcome may have been affected by Cambridge Analytica (CA), a disreputable marketing firm that illegitimately acquired data on millions of Facebook (FB) users and used them to contribute to Donald Trump's campaign. The scandal spurred a vast debate on the looseness of privacy protection in social media and forced FB to promise a drastic reduction of the information released through its Application Programming Interface (API).

Interestingly, the announcement produced the outcry of many social science researchers, who argued that a closure of APIs that does not discriminate between shady marketing initiatives and serious research would harm our capacity to understand online collective phenomena (Bruns *et al.*, 2018; Bastos & Shawn, 2018). An informal survey by Bechmann listed more than 130 "[p]ublications that could not have existed without access to API data" (and the list is not exhaustive).[1] Digital journalism studies are everywhere in this list (with a couple of dozens publications dedicated to online news, three of which published by this very journal) proving the importance of API records for this emergent discipline.

So, in spite of its disturbing political implications (and in part because of them), the CA affair may help to define 'API-based research' as a type of investigation based on the information

[1] https://docs.google.com/document/d/15YKeZFSUc1j03b4lW9YXxGmhYEnFx3TSy68qCrX9BEI/

collected by social media platforms and made available through a set of standardized commands to query, retrieve, filter, format and download records from their databases. In this paper, we discuss this type of research in opposition to CA's marketing operation. We will do so because we believe the nature of this operation has been misunderstood and because we are convinced that digital social sciences are more valuable when defined *in opposition to* the kind of marketing incarnated by CA. Without denying the value of research based on platform APIs, we will defend a type of 'digital fieldwork' that does not depend entirely on them (unlike mainstream 'computational marketing').

Learning from the CA Affair

One good thing about the CA scandal is that it has opened a long overdue discussion on the use of social media data. Yet, this occasion might go wasted if the discussion develops according to the framework proposed by Facebook CEO, Mark Zuckerberg, in his U.S. Senate hearing:

> "It's not enough to just build tools. We need to make sure that they're used for good."
>
> "The issue is that we designed the system in a way that wasn't good. And now, starting in 2014, we have changed the design of the system [so] that it (...) massively restricts the amount of data access that a developer can get." (Associated Press, 2018)

These statements contain two problematic ideas. First, that the data produced by social media are *powerful but also neutral* – extremely useful or extremely dangerous depending on how they are used. Second, that the remedy against misuses is to lock these data *inside* the platforms, so that they do not fall into the wrong hands.

Zuckerberg, to be sure, is not alone in this reasoning. In the press, FB is regularly presented as an adventurer stumbling on the magic lantern of personal data and carelessly setting its genie free. In Zuckerberg's words: "I think it's pretty much impossible… to start a company in your dorm room and then grow it to be at the scale that we're at now without making some mistakes." Social media, we are told, have spawned 'by mistake' avalanches of sensitive data that, if not policed, could influence our votes. Enter CA, which – with its 80 million profiles and its "psychographic analysis" and "behavioural microtargeting" – is held out as the dark pinnacle of this uncanny evolution. But is this an accurate description of the fix we are in?

According to journalistic reconstructions and parliamentary testimony (Cadwalladr, 2018), CA initially tried to buy data from an earlier project at Cambridge University. The *myPersonality* project was based on a FB app delivering a personality quiz taken by 6 million people (Stillwell and Kosinski, 2012). The app also collected the profile records and 'liked' pages, but no information on friends (The Psychometrics Centre Cambridge, 2018). The negotiations broke down because the scholars opposed the commercial reuse of their data. CA then decided to replicate the protocol with the help of Aleksandr Kogan – also a researcher at Cambridge University, who aspired to create his own centre, with students who went on to work for FB (University of Cambridge, 2018).

Kogan's software was more seat-of-the-pants and was completed by fewer than 300,000 people (Kogan, 2018). The 80 million 'stolen' profiles, often mentioned in the media, refer to those 300,000 respondents plus each individual's average number of friends (267). Before FB changed its API in 2015, it was indeed possible to retrieve information not only from the profiles authorising the app, but also from those of their friends. The information harvested from friends was almost as rich (Symeonidis *et al.*, 2015) but crucially did not contain the psychological

information of the personality quiz. 99.5% of the records gathered for CA, therefore, do not contain the information constituting the originality of the firm's 'psychographic approach'.

Additionally, respondents to Kogan's test were recruited through paid services (i.e. *Amazon Mechanical Turk*) (Madrigal, 2018). These microwork platforms are known for 'recruiting' much of their workforce from overseas click farms (Casilli, 2017). Besides being morally questionable, the data provenance raises questions about their reliability and generalizability. Kogan himself, in his U.K. Parliamentary testimony, called the data "noisy" and "worthless" for political advertising (Lomas, 2018). Just using the FB ad platform, he added, would have been more effective, for its greater coverage and finer granularity (Smout & Busvine, 2018). Finally, Kogan misrepresented the purpose of the survey to its respondents, indicating that the data were collected for "research purposes".

So, the 'big data' that CA brags about are questionable in quality and origin, flawed by uneven coverage and unethical in their collection. And, even if they had been 'good data', nothing assures us that they could have any influence on U.S. elections. On its website, CA boasts that its efficacy derives from audience segmentation achieved through a sophisticated personality analysis: "we offer a proven combination of predictive analytics, behavioural sciences, and data-driven ad tech" (Cambridge Analytica, 2018). According to testimony by Christopher Wiley (data scientist at CA), high degrees of 'openness' and 'neuroticism' would, for instance, make a person more receptive to conspiratorial thinking and thus a better target for 'fake news' posts.

As we noted above, however, the company only had psychological data for the 300,000 individuals who were paid to use its app *and not* for their 80 million friends. And, while it may be possible to infer personality traits or by 'homophily' based on the friends who did take the quiz or derive them from the pages liked by the profile (Kosinski *et al.*, 2013), it remains unproven whether such inference can be applied to persuade more effectively than classic political marketing. Even admitting the possibility to infer accurate psychological profiles, it is unclear how personalised contents can be selectively dispatched to them, since FB's advertising platform does not offer tools to target 'states of mind' (García Martínez, 2018). This may be one of the reasons why the 'psychographic' approach did not prove particularly effective when deployed in Ted Cruz's campaign during the primaries (Confessore & Hakim, 2018).

Wiley's testimony, nonetheless, highlights the second arm of CA: the identification of conspiratorial ideas (e.g. the existence of 'deep state' and the solution of 'draining the swamp') that would serve as framing devices for Trump's campaign (Hosenball, 2018). This identification, however, was most probably obtained through surveys and focus groups as much as from online traces. Evidence also exists that CA might have relied on more jarring disinformation techniques. When the British broadcaster Channel 4 (2018) sent out undercover reporters posing as wealthy clients, senior executives at the firm (and its CEO Alexander Nix) focused more on disinformation techniques, bribery, honey traps and kompromat, than on computational prowess, though it is unsure how much of this is also part of the overselling style typical of the company.

In his hearings, however, Zuckerberg did not mention the limited value of CA's data, nor did he point at the minor role that such data have probably played in political advertising. This is hardly surprising, for two reasons. The first is that, in the CA's affair, Facebook conveniently appears as a secondary character, a clumsy but humble infrastructure provider. In fact, most of our critical remarks regarding CA's capabilities to harvest and analyse personal data do not apply to FB itself, which has a monopoly on data access and far greater analytical capabilities. The second reason is that FB sells the same 'computational marketing' peddled by Cambridge Analytics, only on a much larger scale. Both of them sustain their advertising market by boasting of the power of their data and social analytics. This is why the image of social media companies breaching by carelessness the dikes of privacy is misleading. Social media have not heedlessly liberated masses

of pre-existing sensitive data that now need to be bridled. They have purposely and relentlessly built the *self-fulfilling prophecy* of 'computational marketing' and, to do so, created a new type of data devoted to support it.

API-based and digital fieldwork

Phony, flawed and devious as it is, the CA's affair has real consequences. One of the most important is described in the second excerpt from Zuckerberg's testimony: "we have changed the design of the system to… massively restrict the amount of data access". Even though the restriction of API access has begun long before the CA's affair (in 2015 as far as FB is concerned), the increased public concern generated by this and other recent scandals related to online electoral campaigning is likely to increase such closure more.

This is where academia enters the picture. It does so because marketers and developers are not the only ones exploiting the records collected through digital media; academic researchers are also major users of these traces. And for good reasons. Since its inception, social research has been trapped in the divide between rich but limited *qualitative* data and large but crude *quantitative* data. This divide, which seemed insurmountable just a couple of decades ago, is increasingly challenged by researchers working with the vast datasets made available by digital media. For the first time in the history of social sciences, we caught a quick glimpse into what it could mean to trace large collective dynamics down to their single interactions (anonymised_1_1 *et al*, 2015).

Digital records come at a much finer level of aggregation than demography and surveys. They allow not just to gauge categories and opinions, but to examine interactions *word for word*. Before the advent of digital mediation, this was only possible for the small and situated communities that researchers could observe ethnographically. Nowadays, a similarly sensitive inquiry is possible for larger populations, stretching farther in time and space. This type of quali-quantitative investigation is particularly crucial in digital journalism studies, because of the way in which information travels in the contemporary media system. In the past, the circulation of news was relatively stable and, to a large extent, determined by the form of the infrastructure dedicated to the dissemination of newspapers and television and radio signals. With the advent of the digital networks, news has started to spread in more complex ways and to mutate while spreading – a type of diffusion that is both crucial to investigate and impossible to follow without the help of digital traces (anonymised_1, 2018). No wonder that researchers are shouting out loud now that marketeers' folly offers an excuse for platforms and institutions to lock away their data – not deleting them (mind the difference) – just making them inaccessible to (or more expensive for) out-of-house researchers.

Yet, if it is fair to complain when academic projects are treated in the same manner as marketing campaigns, academics may deserve part of the blame. While the use of the new data made available by digital media may allow to overcome the quali-quantitative divide, it also exposes social researchers to the bias of the infrastructures allowing the collection of digital records. This risk is especially high for *API-research*, which we define as

> *an approach to computational social sciences and digital sociology based on the extraction of records from the datasets made available by online platforms through their application programming interfaces.*

Compared to previous techniques for collecting digital records, social media's APIs arrived as a godsend, offering huge quantities of data accessible in a few clicks and ready to be analysed (at least in comparison to much messier inscriptions that one could scrape directly form the Web). Striving to impose themselves and to sell their advertising space, social media have concentrated

a great deal of online discussions in a few "platforms" (Gillespie, 2010) and made *some* of their traces accessible through normalised 'firehoses'.

By building the infrastructures necessary to support and trace a growing amount of online interactions and by making the resulting records available through their APIs, platforms have significantly reduced the costs of social data (or rather accepted to pay a large part of such costs). In exchange they obtained and promoted a dependency on their infrastructures and through them 'captured' a growing share of academic research and commercial marketing (Nechushtai, 2018). The facility of API-research came at the price of accepting the particular standardisation operated by social media platforms and the bias that comes with it (Puschmann & Burgess, 2013). A bias to which we, researchers in "digital sociology" (Marres, 2017) and "computational social sciences" (Lazer *et al.*, 2009), have often turned a blind eye. In a consumerist frenzy, we stockpiled data as mass-produced commodities – literally, if you think that a lot of social media data derive from the exploitation of overseas click farms and industrial techniques (e.g. the so-called 'fame bots') to boost likes, followers and other counts (New York Times, 2018). API-based research is guilty (in part at least) of spreading the hype of social media data, in reducing the diversity of digital methods to the study of online platforms, and in spreading the received ideas that Facebook, Google, Twitter and their likes are the masters of online debate and that there are no alternatives to living on the breadcrumbs of their APIs.

The closure of social APIs should remind us that rich collective dynamics existed online long before social platforms (and still exist alongside them). People were finding information on the Web before Google and passing it along before Twitter. They were selling online before Amazon and networking before Facebook (Rheingold, 1993). And all the while, there were researchers and journalists investigating the digital records of these interactions. APIs are not the only way to study online phenomena. Crawling and scraping but also clicking and reading as well as collaborating with the websites generating the records remain valid alternatives. To be sure, these research methods have their own biases and limitations, being in general more difficult and time-consuming and producing records that are more heterogenous and noisy. However, these forms of collection are, we believe, more than a 'necessary evil', and if carried out conscientiously offer three advantages.

First, while the APIs detach the digital inscriptions from the context of their production ('application programming interfaces' are, by definition, different from 'user interfaces'), more direct forms of harvesting force researchers to observe online dynamics through the same interfaces as the actors they study. Some forms of collection even require the collaboration of the users generating the records or of the administrators managing them (e.g., when the archive of a discussion group is made available by its moderators) and thus encourages a greater dialogue with online actors.

Second, the closure of social APIs encourages researchers to diversify their 'informational diet'. Offering large amounts of cheap and standardized data, mainstream platforms attracted most of our attention. Yet, smaller sources exist that are just as interesting if not more, depending on the topic of investigation (anonymised_1 *et al*, 2014). We relied on Facebook, Twitter, Google etc. not because their data were finest or the fittest, but because they were the most visible as well as the most easily available. Now that these datasets are becoming harder to access, incentives grow to consider other sources.

Third, even for researchers focusing on mainstream platforms, standard APIs are not the only solution. Access can be granted by industry-research partnerships (King & Persily, 2018) such as the one experimented by *Social Science One*. This initiative seeks to associate scholars and FB's engineers to provide datasets especially curated for electoral research - interestingly with explicit reference to the CA affair opening this paper. "Funding partners recognize the threat presented

by the misuse of FB data, including by an academic associated with Cambridge Analytica. That is why the data made available for research by Social Science One is first reviewed and cleared by our commission members, external privacy experts, and FB's privacy and research review teams"[2]. It remains to be seen how the curation of the data affects the types of critical research to be undertaken.

API restrictions may end up being a good thing if they encourage researchers to return to fieldwork. By 'digital fieldwork', to be sure, we do not refer exclusively to close reading and ethnographic methods. Reducing our reliance on standard API data does not mean giving up the project of harvesting rich data on large populations, but it does imply investing in the efforts necessary to cultivate such data. API querying itself can be a form of digital fieldwork when it is not a wholesale accumulation of big-for-the-sake-of-big data, but a careful work of extraction carried out in collaboration with the platforms and their users.

A few political implications

Rediscovering the value of digital fieldwork is crucial to counteract the tendency to reduce online collective life (and collective life *tout court*) to the records left in social media. We touch here upon the most important fallacy of Zuckerberg's hearing and of the whole debate about CA: the idea that social platforms represent the only or the best source of information on social phenomena. This idea, we believe, is methodologically misleading and politically dangerous.

There is certainly a lesson to be learned from the CA story, but it is not that our lives are monitored by powerful political actors (which we should also worry about, but not in this case). Rather, the affair makes painfully clear how easily public debate can be polluted by computational marketing. It reminds us that the influence of social media data and analytics is greater when it serves the purpose for which it been developed in the first place – which is promoting the kind of superficial attention most suited to the contemporary the system of advertisement and entertainment (Jenkins *et al*, 2013; anonymised_1, 2019).

Without a sufficient injection of reflexivity and *re*purposing (anonymised_2, 2013) social media data are partial to some conversations rather than others. Similar to the way in which other forms of algorithmic intelligence (e.g., predictive policing, (Benbouzid, 2015)) are fraught with danger not because they succeed in their fine-grained social control, but because they fail in the way that it is biased against certain groups, so standard social media analytics are partial to certain types of public debate. These analytics have been developed to promote an audience of consumers whose features are opposite to those of a healthy democratic public. Because they focus on effortless actions such as click, like and share, social analytics promote a type of engagement that is light and short-lived; because they measure these actions in an individualized (rather than communitarian) way, they foster individual popularity rather than collective action (anonymised_2, 2018). This is why resisting to be 'captured' by the infrastructure of social media (Nechushtai, 2018) is crucial not only methodologically, but also politically. The more public opinion and political actors believe that elections can be decided by social media data and computational marketing and invest in them instead of promoting richer political discussions, the worse will be the quality of public debate.

This is why social media data can never be used naively and without reflection on their context of production and their biases (anonymised_1 *et al*, 2018). Research *through* media platforms should always be also research *about* media platforms. Such a reflexive posture is particularly important for digital journalism studies, whose crucial political mission is to investigate the very fragile

[2] https://socialscience.one/our-facebook-partnership

equilibriums that allow the survival of a functioning public debate – even in the age of social media (Gray *et al*, forthcoming). By this we do not want to diminish the achievements accomplished *through* (and not *by*) social media – for instance, the way in which activists of all causes have repurposed platforms to disseminate suppressed information or mobilise public opinion (Benkler, 2006; Gerbaudo, 2012). But all these things are accomplished by individuals and groups, rather than by data and algorithms, and are poorly captured by standard social analytics and API data. To understand and support the work of the social and political actors who strive to *re*purpose online platforms and make them compatible with a healthy public debate, we should thereby renounce some of the comforts of the API and return to our digital fieldwork.

References

Associated Press. 2018. "US Facebook Zuckerberg." *AP Archives*. http://www.aparchive.com/metadata/youtube/696c11b65984354a350b4e8139e385b2.

Bastos, Marco, and Shawn T. Walker. 2018. "Facebook's Data Lockdown Is a Disaster for Academic Researchers." *The Conversation*, April 11. https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533.

Benbouzid, Bilel. 2015. "From Situational Crime Prevention to Predictive Policing. Sociology of an Ignored Controversy." *Champ Pénal/Penal Field,* VII (July): 1–19. doi:10.4000/champpenal.9066.

Benkler, Yochai. 2006. *The Wealth of Networks. How Social Production Transforms Markets and Freedom*. New Haven: Yale University Press. http://www.benkler.org/wonchapters.html.

Bruns, Axel. 2018. "Facebook Shuts the Gate after the Horse Has Bolted, and Hurts Real Research in the Process." *Internet Policy Review*, April 25. https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786.

Cadwalladr, Carole. 2018. "'I Made Steve Bannon's Psychological Warfare Tool': Meet the Data War Whistleblower." *The Guardian*, March. https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump.

CA. 2018. "The CA Advantage." *CA Website*. https://ca-political.com/ca-advantage.

Casilli, Antonio. 2017. "Digital Labor Studies Go Global: Toward a Digital Decolonial Turn." *International Journal of Communication* 11 (Special section "Global Digital Culture"): 3934–3954. https://hal.archives-ouvertes.fr/hal-01625688/.

Channel 4 News. 2018. "Revealed: Trump's Election Consultants Filmed Saying They Use Bribes and Sex Workers to Entrap Politicians." *Channel 4*, March 19. https://www.channel4.com/news/cambridge-analytica-revealed-trumps-election-consultants-filmed-saying-they-use-bribes-and-sex-workers-to-entrap-politicians-investigation.

Confessore, Nicholas, and Danny Hakim. 2018. "Data Firm Says 'Secret Sauce' Aided Trump; Many Scoff." *The New York Times*, March 6. https://www.nytimes.com/2017/03/06/us/politics/cambridge-analytica.html.

García Martínez, Antonio. "The Noisy Fallacies of Psychographic Targeting." *Wired*, March. https://www.wired.com/story/the-noisy-fallacies-of-psychographic-targeting.

Gerbaudo, Paolo. 2012. *Tweets and the Streets: Social Media and Contemporary Activism*. London: Pluto Books.

Gillespie, Tarleton. 2010. "The Politics of 'Platforms.'" *New Media & Society* 12 (3): 347–64. doi:10.1177/1461444809342738.

Gray, Jonathan, Liliana Bounegru, and Tommaso Venturini. 2018. "The Infrastructural Uncanny: 'Fake News' and the Fabric of the Web as Matter of Concern." *New Media & Society*, no. forthcoming.

Hosenball, Mark. 2018. "Ex-Trump Aide Bannon Promoted 'culture War' - CA Whistleblower." *Reuters.Com*, May 17. https://af.reuters.com/article/worldNews/idAFKCN1IH370.

Jenkins, Henry, Sam Ford, and Joshua Benjamin Green. 2013. *Spreadable Media*. New York: New York University Press. doi:10.1017/CBO9781107415324.004.

King, Gary, and Nathaniel Persily. 2018. "A New Model for Industry-Academic Partnerships." Cambridge Mass. https://gking.harvard.edu/partnerships.

Kogan, Aleksandr. 2018. "Written Evidence Submitted by Aleksandr Kogan." *Text Submitted to the The Digital, Culture, Media and Sport Committee of the UK Parliament*.

https://www.parliament.uk/documents/commons-committees/culture-media-and-sport/Written-evidence-Aleksandr-Kogan.pdf.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” *Proceedings of the National Academy of Sciences* 110 (15): 5802–5. doi:10.1073/pnas.1218772110.

Latour, Bruno, Pablo Jensen, Tommaso Venturini, Sébastian Grauwin, and Dominique Boullier. 2012. “‘The Whole Is Always Smaller than Its Parts’: A Digital Test of Gabriel Tardes’ Monads.” *The British Journal of Sociology* 63 (4): 590–615. doi:10.1111/j.1468-4446.2012.01428.x.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al. 2009. “Computational Social Science.” *Science (New York, N.Y.)* 323 (5915): 721–23. doi:10.1126/science.1167742.

Lomas, Natasha. 2018. “Kogan: ‘I Don’t Think Facebook Has a Developer Policy That Is Valid’.” *TechCrunch*, April 24. https://techcrunch.com/2018/04/24/kogan-i-dont-think-facebook-has-a-developer-policy-that-is-valid/?guccounter=1.

Madrigal, Alexis C. 2018. “What Took Facebook So Long?” *The Atlantic*, March 18. https://www.theatlantic.com/technology/archive/2018/03/facebook-cambridge-analytica/555866/.

Marres, Noortje. 2017. *Digital Sociology: The Reinvention of Social Research*. Maden: Polity Press.

Metz, Rachel. 2018. “The Scientist Who Gave CA Its Facebook Data Got Lousy Reviews Online.” *MIT Technology Review (The Download)*, March 21. https://www.technologyreview.com/the-download/610598/the-scientist-who-gave-cambridge-analytica-its-facebook-data-got-lousy-reviews/.

Nechushtai, Efrat. 2018. “Could Digital Platforms Capture the Media through Infrastructure?” *Journalism* 19 (8): 1043–58. doi:10.1177/1464884917725163.

Puschmann, Cornelius, and Burgess, Jean. 2013. “The Politics of Twitter Data.” *SSRN Electronic Journal*. doi:10.2139/ssrn.2206225.

Rheingold, Howard. 1993. *The Virtual Community: Homesteading on the Electronic Frontier*. Cambridge Mass.: MIT press.

Smout, Alistair, and Douglas Busvine. 2018. “Researcher in Facebook Scandal Says: My Work Was Worthless to CA.” *Reuters*. https://www.reuters.com/article/us-facebook-privacy-cambridge-analytica/researcher-in-facebook-scandal-says-my-work-was-worthless-to-cambridge-analytica-idUSKBN1HV17M.

Stillwell, David J, and Michal Kosinski. 2012. “MyPersonality Project : Example of Successful Utilization of Online Social Networks for Large-Scale Social Research.” *The Psychometric Centre, University of Cambridge*.

The New York Times. 2018. “The Follower Factory.” *The New York Times*. https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html.

The Psychometrics Centre Cambridge Judge Business School. 2018. “Statement on Politics.” https://www.psychometrics.cam.ac.uk/about-us/media-reports/statement-on-politics.

University of Cambridge. 2018. “Statement from the University of Cambridge about Dr Aleksandr Kogan.” https://www.cam.ac.uk/notices/news/statement-from-the-university-of-cambridge-about-dr-aleksandr-kogan.

Youyou, Wu, Michal Kosinski, and David Stillwell. 2015. “Computer-Based Personality Judgments Are More Accurate than Those Made by Humans.” *Proceedings of the National Academy of Sciences* 112 (4): 1036–40. doi:10.1073/pnas.1418680112.