Prized Assets for Web Tracking

Sylvia E. Peacock
York University
Dept. of Business and Society
Social Science
Toronto, ON, Canada
+1-416-462-3454
speacock@yorku.ca

ABSTRACT

In this study, I pinpoint the group of online Canadians who are predominantly affected by online personal data extractions, sometimes called web tracking. Due to the lack of regulations and online information market failure, online personal data extraction is ubiquitous. Little knowledge exists about who the users are that are mainly affected by web tracking. Two large population surveys, called the Canadian Internet User Survey 2010 and 2012, offer comprehensive and comparable measures to gauge the intensity and extent of people's online activities. An Internet intensity index is proposed to measure Canadian online activity scaled between zero (offline) and 100 (intensely active Internet user). Results indicate that the data of mature, well-off and educated men and women are captured by the web tracking technologies of online companies, to an increasing degree. Other population segments appear less affected. From a marketing perspective, the targeted group is the most coveted of all consumer groups. Thus, corporate self-regulation seems unlikely in the light of the current results. The overuse of social media and the Internet commons for the purpose of personal user data extraction might lead to an underuse of the Internet's possibilities to connect via social media. Today a market-led approach that essentially leaves web tracking unregulated may present first indications of online companies steering us towards a virtual *tragedy of the commons*.

Categories and Subject Descriptors

Human-centered computing—User studies, Applied computing—Economics, Applied computing—Sociology, Applied computing—Electronic data interchange, Social and professional topics—Consumer products policy, Computing methodologies—Hybrid symbolic-numeric methods.

General Terms

Economics, Human Factors, Theory, Legal Aspects.

Keywords

Online User Agency; Social Media; Web Tracking; Age; Education; Income; Big Social Data.

1. INTRODUCTION

One of the most puzzling aspects of studying the effects of big

social data on user agency, is that little is known about who the most intense Internet users are. There are Internet users who use email but do not spend much time online. There are also those who spend hours on social media, gaming platforms, online shopping, and live more of their lives online rather than offline. Research regarding who *accesses* the Internet exists [13], particularly after the release of the Information and Communication Technology (ICT) development index [10]. Ever since the development of this index, studies on ICT access have spawned new internationally comparative research [3,6,11,17] and a body of index assessment research [14,19]. For Canada, the online population has been identified as more than 80% [20,21]. But the discrepancy between Internet access and the intensity and extent of Internet use is a gap this paper seeks to address. An understanding of who these Canadians are may advance a more analytical insight of what today's web tracking means for the online user agency of intense Internet users.

2. BACKGROUND

2.1 Online User Agency

An important assumption about user agency is that people who use technology prefer to be the experts of their own situation [4]. This supposition includes online users who voluntarily agree to have their personal data extracted in order to be targeted by advertisements that fit into their consumer profile [22]. In a recent literature review, Acquisti, Brandimarte and Loewenstein point out that people show an overwhelming preference for making their own personal decisions about the destinations of their personal data, even if some behavioral choices indicate otherwise [1]. They explain the inconsistency between online behavior and offline choices in conceptual frameworks of uncertainty, context-dependence, and malleability that online users face [1]. Whatever the explanation might be, online actors are left with few options regarding the use of their online personal data. Users either 'agree' to the extraction of their personal data and non-negotiable terms and conditions, or must go offline, thus 'exit' without the option to 'voice' their own preferences.

2.2 Public Sphere versus Market Exchange

Currently, two dominant opinions prevail as to what the Internet should represent: a public sphere or a market place [18]. On the one hand there are those who believe it is a public (anonymous) meeting place, open to all who choose to come and exchange ideas and opinions. Habermas coined such a meeting place a 'public sphere' [9]. On the other hand there is a belief that the Internet is predominantly a place where people exchange goods and services and engage in marketing activities. This opinion implies that all online activity comes at a price. CEO's and companies such as Jeff Bezos of Amazon.com, Zuckerberg of Facebook, Eric Schmidt of Google, or the Waltons of Walmart have been moving the Internet in a direction of online market exchanges.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

SM Society '15, July 27-29, 2015, Toronto, ON, Canada

ACM 978-1-4503-3923-0/15/07.

<http://dx.doi.org/10.1145/2789187.2789199>

Some online actors may be indifferent as to what the ‘ideal’ Internet ought to represent as long as they are able to receive what they are seeking. The introduction of a price to the exchange of information and goods changes the incentives for online participants. Much of the price paid on the Internet is not in the local currencies but with personal data, web tracks that people leave behind when going online. With the Internet moving towards a market place, an increasing number of interactions are driven towards increasingly profitable data extractions [18].

2.3 Social Media and Applications

Important applications that use a market led approach are social media, like twitter or Instagram, or web search engines like Google, Bing or Yahoo!. Of course, sellers of online services and goods like Amazon.com also use people’s personal data in exchange for their website access to increase their profits from online activities and predict purchasing behavior. There are myriad other applications, like online smut and bit torrents which make additional profits by redirecting web traffic and injecting spyware, adware or malware [8]. Discoveries of factory installed crapware on user computers are numerous [2,8], with a recent one being an executable file called superfish, that logs a user’s online behavior in a file and read by the firms remote server and sells the data to market analysts [12]. In this instance, crapware is installed to lower the price for cheap consumer laptops and still maintain a profit margin [2]. The list could go on, but it suffices to assume that an increase in online activities and more access to a variety of websites goes along with a higher exposure to personal data extractions [5,12]. In other words, *the longer, more often, and varied websites an online user accesses on the Internet, the more she is exposed to surreptitious personal data extractions.*

2.4 Why Canada?

When investigating online personal data extraction in a networked world, it appears to be irrelevant where one lives. But penalties and incentives matter when unfair market exchanges are pursued by quasi-monopolies.¹ The free-market data management regime of the U.S., which is characterized by self-regulation and little institutional oversight, is currently setting an international precedence. Canada, as her nearest neighbor and with most of the online users using English language websites, is most intimately affected by U.S. Internet policies.

Thus, a study of the Canadian case exemplifies the effects of U.S. policies on international third parties. It allows us to make inferences about the uniqueness and size of the Canadian subpopulation and the varying degrees to which they are submitted to web tracking technologies.

3. DATA & METHODS

The empirical results are based on data from the Canadian Internet Use Survey (CIUS), a large population survey conducted as a supplement of the Labour Force Survey [20,21]. This representative Canadian survey is based on approximately 30.000 Canadian households, with a response rate of 81% for the household component and 69% for the individual component. The datasets represent roughly 27 million weighted respondents (see <http://bit.ly/1CQLERC>). Both surveys, the 2010 and 2012 CIUS, were designed to measure the extent and scope of online activities of the Canadian population.

¹ For example, Google Inc. is in the middle of an anti-trust investigation in Europe, see <http://bit.ly/1DkYYI4>.

3.1 Canadians – Online and Offline

For both datasets, Table 1 shows the most important descriptive measures used in multivariate model (for more information, please download files at www.researchgate.net/Sylvia_Peacock). In short period of time, thus between 2010 and 2012 general population distribution characteristics scarcely change. Most people in Canada live either in a metropolitan or an urban area (together 78%), with an average age in the low 40s, a college degree, about two thirds of the population in paid employment (with some variations throughout the year, of course), an almost equal number of men and women in the population, and an average income in the third quintile (\$41.001 – \$65.000).

Table 1. Canadian Demographics (all %)

	$\Sigma^{(a)}$	Intense Surfers ^(b)		Offliners	
Demographics	2010/12	2010	2012	2010	2012
Geo-location	%	%	%	%	%
Montreal	12	17	12	12	11
Toronto	17	17	24	15	13
Vancouver	7	18	10	5	6
Urban Area	42	43	45	39	40
Rural Area	22	4	9	29	30
Age Group					
16-24 Years	15	52	25	2	1
25-34	17	34	36	5	3
35-44	17	12	21	8	6
45-54	19	13	13	15	15
55-64	15	4	4	21	19
65+ Years	17	0	0	50	55
Men	49	78	63	47	49
Education					
High School	37	42	16	65	68
College	41	21	41	29	27
University	22	37	42	6	5
Student	12	51	24	1	1
Main Activity					
Paid Work	63	47	81	32	31
Unemployed	4	8	7	6	2
At Home	32	45	12	65	67
Income (\$)					
Q1-25.000	15	12	12	31	33
Q2 → 41.000	18	5	11	31	30
Q3 → 65.000	20	41	19	18	19
Q4 → 100.000	21	14	24	13	11
Q5 +100.000	25	27	35	7	7
Weighted N, in million	28 / 27	0.1	2.3	5.5	4.8

Source: [20, 21].

(a) No significant changes between 2010 and 2012, (b) ‘Intense Surfers’ are online users who score 80+ on the I3.

Some of the more interesting developments between 2010 and 2012 include changes to the socio-economic background of intense Internet users. Note the higher median age, educational and income level of this population segment. From an online marketing perspective, this is an indication of the unequal distribution of web targeting practices towards a particularly precious group of people.

3.2 The Internet Intensity Index (I3)

The CIUS covers a range of 22 different websites that respondents may have accessed in the last year. These answers are added up to range between 0 and 22. Additionally, the average weekly hours spend accessing the Internet were included in the questionnaire (*hcat*). The first as an additive and the second as a weighted measure were combined and standardized on an Internet Intensity Index (*I3*) scale ranging from 0 to 100, with the former indicating offliners and the latter forming an index that makes the intensity of Internet use comparable:

$$I3 = \left[\sum_{22}^1 x * hcat \right] * \frac{100}{\max(x)}$$

Cronbach's alpha was 0.92 for 2010 and 2012, indicating a fairly good scalability of the 22 items on the questionnaire. A varimax rotated factor analysis revealed between two (2010) and six (2012) underlying factors when all the standardized website items were entered into an analysis. As questions were asked in exactly the same fashion in both survey years, it remains an open question why underlying factors multiplied between 2010 and 2012. Due to the coding of the Internet Intensity Index, the variable can be roughly understood as a percentage, with offliners coded '0' while people using the entire range of websites on the Internet, e.g., social media, online banking, travel agencies, or gaming for several hours every day are up in the last quintile of the index (between 80 and 100). Table 2 shows the means and standard deviation for the index in 2010 and 2012.

Table 2. Means and Standard Deviation for the I3

Internet Intensity Index	Scale	2010 (Std. Dev.)	2012 (Std. Dev.)
Canadians	0-100	14.2 (15.7)	42.1 (27.1)
Weighted N (mill.)		27.4	28.1

Source: [20,21].

On average, the intensity of Internet use has significantly increased between 2010 and 2012. While average intensity was low with a small standard deviation in 2010, the distribution has flattened with a triple-fold average increase in 2012. The last problem this work needs to address is whether the discovered upward trends in user intensity and selectivity of the online group are not mere artifacts due to steady technology diffusion. Therefore, the two datasets are combined and significant changes investigated between 2010 and 2012 while controlling for socioeconomic population factors.

3.3 Multivariate Results

In Table 3, the results for 2012 show significant changes in comparison to outcomes in 2010; thus, the effects are additive. Non-significant results are noted as 'ns'. All covariates are dummies, coded 0/1, except for the dependent variable, the Internet Intensity Index.²

² For example, let us assume we want to calculate from the results in Table 3 with which intensity a young graduate student in Toronto (for instance, a woman, 25 years old) uses the Internet in 2012, while living in a household with approximately \$30,000, working part-time. The coefficients add up to the following result: (3.8+2.9) + 2.9 + (12.3+20.5) + (7.2+11.1) + (5.4+2) + (-3.3) + 4.0 = 68.8. Thus in 2012 this fictitious person

If anything, the covariates for the geo-location indicate that urban Canada has increased the extent of Internet use between 2010 and 2012, compared to rural Canada (including PEI). In general, more intense Internet users live in cities, but there is no widening gap to rural Canadians.

The largest changes between 2010 and 2012 towards more intense Internet use happen in the groups of older Canadians, better educated, and wealthier people, or all of these together. Sex differences virtually disappear during this time.

Table 3. Significant Changes in the I3

Covariates ^{a)}	2010	Δ 2012
Geo-location		
Montreal	1.6	ns
Toronto	2.9	ns
Vancouver	4.2	ns
Urban Area	2.1	1.27
Rural Area (cg) ^{b)}	--	--
Age Group		
16-24 Years	14.4	21.5
25-34	12.3	20.5
35-44	7.4	18.3
45-54	4.1	12.2
55-64	2.4	7.8
65+ Years (cg)	--	--
Gender		
Women (cg)	--	--
Men	1.2	-1.7
Education		
High School (cg)	--	--
College	3.1	6.2
University	7.2	11.1
Currently in Education		
No Student (cg)	--	--
Student	5.4	2.0
Main Activity		
Unemployed (cg)	--	--
Paid Work	-3.3	ns
At Home	-4.7	ns
HH Income (\$)		
Q1-25.000 (cg)	--	--
Q2 26 → 41.000	ns	4.0
Q3 42 → 65.000	1.2	6.8
Q4 66 → 100.000	1.4	9.5
Q5 +100.000	2.4	10.2
II Index (constant)	3.8	2.9
N	22623	22615

Source: [20,21].

a) Coefficients p<0.01; b) Control group (cg).

Thus, what do these results mean for the initial question of this study? Remember that at the onset the idea was to get a more solid understanding of the characteristics of people that use the Internet most intensely and therefore have the highest propensity of being web tracked. In Table 3, groups with higher (and increasing) scores relentlessly have their personal data extracted and stored in big data bases, their data entries retrieved, analysed and their online activities targeted by behavioral advertising.

has an *I3* of almost 70 points, situating her almost one std. dev. removed from the mean (42 points, 27 points Std. Dev.).

In the wake of the current results it is sufficiently clear, that the most intense Internet users coincide with what seem to be the most prized assets for online companies using web tracking, behaviorally targeted advertising, or real-time bidding. The higher income, middle-aged, well-educated and most often urban and metropolitan dwellers represent one of the most coveted targeted marketing groups in the industry. This group contains the most intense Internet users, as table 3 shows. If nothing else, the current model underlines why most attempts to politically interfere in today's surreptitious online personal data extraction are stonewalled by the online industry [2,12].

4. CONCLUSIONS

For the purpose of online marketing, Internet users are submitted to often surreptitious personal data extractions on the Internet, via social media platforms, search engines, browser tracking, flash cookies, or factory-installed crapware on newly bought devices. Online personal data market exchanges – even if for online users these are more akin *unconscionable contracts* than actual exchanges – are a dominant feature for today's online platform users [18]. It seems widely agreed upon that the built-in trade-off between offering online content for the price of extracting, storing, and analysing personal data, the so-called “convenience versus privacy” debate, is resolved in favour of the former [7: 217f.]. From an online marketing perspective few incentives exist for self-regulation, because the most active Internet users are the most coveted marketing group. Any self-restriction would immediately translate into a severe competitive disadvantage, present or future, in an unregulated online information market [16: 28f.].

Intense Internet users are the ones who are for the most part exposed to a now common but largely unspecified personal data extraction while offliners remain untouched [15]. These conclusions are of particular interest regarding a population segment that is rarely covered in population surveys: Teenagers and children, though they are increasingly prone to use social media. Young people are usually excluded from population surveys because research guidelines and ethical considerations are much more stringent for underage respondents. However, these guidelines and ethical considerations do not apply to web tracking and behaviorally targeted advertising using the personal online data of underage Internet users.

A missing link into an inquiry on the effects of web tracking has thus far been an understanding of the intensity of people's Internet use and the characteristics of the most active online population. Due to the results shown in this paper we know who these people are and what the trends are: It are mainly the personal data of middle-aged, well-educated, wealthier segments of the Canadian population that are tracked and with high probability analysed for marketing purposes. Usually Internet users expect behaviorally targeted advertising when using social media, but the stored data tracks go beyond the uploading of photos and text entries. They encompass the machine environment, login duration, idling, and webvisits during social media use (and sometimes long after, if the user chooses to remain to login). The trouble with this is that it turns the Internet into less of a public sphere. Increased web tracking is morphing into the Internet into an online sphere ‘by the corporation, of the corporation, and for the corporation’, where the interests of the people clearly are not at center stage.

While circumventing technological determinism, it remains to be seen whether the socially undesirable outcomes of web tracking and behaviorally targeted advertisements will increase. On the one

hand one might argue that whoever decides to use social media has nobody but herself blame for offering her personal data to online companies. But this approach assumes adequate online agency, sufficient information, an ability to refuse to pay up, and last but not least, an easy substitution of online goods and services that extract too many personal data by others that offer a better deal [16]. Indeed, for online access to retailers on the Internet or grocery stores, it may be the substitutable; however, there are no alternatives to Twitter, Facebook, or Google [18].

The next puzzle that needs to be tackled is a possible *tragedy of the online commons*. In due time, the overuse of the Internet commons, and in particular social media, for the purpose of personal user data extraction might lead to an underuse of its connective possibilities, as online users become increasingly disenchanting. Perhaps online companies are not just distributing personal data extraction unequally, but also are extracting more than their fair share of the available profits from an unregulated Internet commons – if the assumption is correct that the Internet is not just a market place but a public sphere, too.

5. REFERENCES

- [1] Acquisti, A., Brandimarte, L., and Loewenstein, G. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221, 509-514. DOI: <http://dx.doi.org/10.1126/science.aaa1465>.
- [2] Arthur, C. 2015. Lenovo demonstrates that malware is big business. *The Guardian (online)*, February 20, 2015, <http://gu.com/p/462xy/stw>.
- [3] Ayanso, A., Cho, D.I., and Lertwachara, K. 2014. Information and communications technology development and the digital divide: A global and regional assessment. *Information Technology for Development* 20, 1, 60-77. DOI= <http://dx.doi.org/10.1080/02681102.2013.797378>.
- [4] Benello, C.G. 1981. Technology and Power: Technique as a Mode of Understanding Modernity. In: Christians CG and van Hook JM (eds.) *Jacques Ellul: Interpretive Essays*. University of Illinois Press: Urbana, 91-107.
- [5] Christl, Wolfie. 2014. *Kommerzielle digitale Überwachung im Alltag*. Cracked Labs Institut für digitale kritische Kultur, report online available <http://bit.ly/1tyedPR> (accessed March 2015).
- [6] Dobrota, M., Jeremic, V., and Markovic, A. 2012. A new perspective on the ICT development index. *Information Development* 28, 4, 271-280. DOI= <http://dx.doi.org/10.1177/02666669/2446497>.
- [7] Dumas, M.B. 2012. *Diving into the Bitstream: Information Technology meets Society in a Digital World*. New York: Routledge.
- [8] Edelman, B. n.d. *Spyware, Adware, and Malware: Research, Testing, Legislation, and Suits*. Available at: <http://www.benedelman.org/spyware/> (accessed March 2015).
- [9] Habermas, J. 1989. *The Structural Change of the Public Sphere*. Cambridge: MIT Press.
- [10] ITU. 2009. New ITU ICT Development Index compares 154 countries: Northern Europe tops ICT developments. Available at: http://www.itu.int/newsroom/press_releases/2009/07.html.

- [11] James, J. 2012. The ICT development index and the digital divide: How are they related? *Technological Forecasting & Social Change* 79, 567-504. DOI= <http://dx.doi.org/10.1016/j.techfore.2011.06.010>.
- [12] Kleinz, T. 2015. Das Geschäft mit der Crapware. *Die Zeit*. Feb. 24, 2015. Available at: <http://bit.ly/1MhHJU0>.
- [13] Korupp, S.E. 2006. No Man is an Island: The Influence of Knowledge, Household Settings, and Social Context on Private Computer Use. *International Journal of Internet Science* 1,1, 45-57.
- [14] Kyriakidou, V., Michalakelis, C., and Sphicopoulos, T. 2013. Assessment of information and communication technology maturity level. *Telecommunication Policy* 37, 48-67. DOI = <http://dx.doi.org/10.1016/j.telpol.2012.08.001>.
- [15] Landau, S. 2015. Control use of data to protect privacy. *Science* 347, 6221, 504-506. DOI: <http://dx.doi.org/10.1126/science.aaa4961>.
- [16] McChesney, R.W. 2013. *Digital Disconnect: How Capitalism is turning the Internet against Democracy*. The New Press: New York, London.
- [17] Miranda, L.C.M., and Lima, C.A.S. 2012. Trends and cycles of Internet evolution and worldwide impacts. *Technological Forecasting & Social Change* 79, 744-765. DOI= <http://dx.doi.org/10.1016/j.techfore.2011.09.001>.
- [18] Peacock, S.E. 2014. How web tracking changes user agency in the age of Big Data: The used user. *Big Data & Society*. DOI= <http://dx.doi.org/10.1177/2053951714564228>.
- [19] Song, J., and Wang, E. 2013. China's Information and Communication Technology in Geographic Perspective. *Eurasian Geography and Economics* 53,4, 502-526. DOI= <http://dx.doi.org/10.2747/1539-7216.53.4.502>.
- [20] Statistics Canada. 2011. *Canadian Internet Use Survey 2010*. Public use microdata file, available on request.
- [21] Statistics Canada. 2013. *Canadian Internet Use Survey 2012*. Public use microdata file, available on request.
- [22] Turow, J. 2012. The Disconnect About What People Say and Do About Privacy. *Journal of Law* 2, 479-482.