Summer September 8, 2013

The Privacy Problem in Big Data Applications: An Empirical Study on Facebook

Jerzy Surma, *Warsaw School of Economics*

The Privacy Problem in Big Data Applications: An Empirical Study on Facebook

Jerzy Surma
Warsaw School of Economics
Warsaw, Poland
E-mail: jerzy.surma@sgh.waw.pl

***Abstract*—When using mobile phones, credit cards, electronic mail, browsing social networks etc., contemporary consumers leave behind thousands of digital footprints. Each footprint reflects actual actions that we take in given place and time. The analysis of thousands of such footprints conducted among large groups of people allows us to examine human behaviour on a scale that has never been imagined in scientific studies concerning psychology and sociology. The results of those analyses already have a significant influence on contemporary management, especially when it comes to new business opportunities in companies that employ business models based on the one-to-one relations with their customers. Nevertheless, this outstanding opportunity implies an enormous privacy problem. We will illustrate this issue by an empirical research based on the data gathered from Facebook, where users are using privacy controls that allow displaying their content only to a selected group of people. Users of such controls will likely continue positing more, even as their network grows or becomes sparser. We test these predictions using a dataset from Facebook gathered from a sample of college students and find statistical support for them. Our conclusions are that individuals are relatively prudent and are actually very aware of the social norms.**

Keywords: big data analysis, social netwrok sites, privacy problem

I. Introduction

The analysis of digital footprints (credit card transactions, mobile phone calls, e-mails replays, RFID mass transit cards, movements captured by GPS in smart devices, blog post, “likes” in social networks sites, etc.) that reflect human behaviour is a particularly important factor that stimulates paradigm shifts with in the business management, which is focusing on extensive data-based decisions making by employing detailed and in-depth information on customers. Access to extensive, real-time volumes of behavioural customer’s data, with special regard to their social relationships, becomes a basis for the development of advanced analytical customer management systems. One of their most important elements are social relationships and interactions between customers associated with integration of social networks sites with a corporate information systems. The ongoing research seems to support the hypothesis by Petland [1] on the so called ‘honest signals’, which may be paraphrased as follows:” the analysis of digital footprints, reflecting actual behaviour of people, allows us to foresee their future actions, providing data that may be used for business purposes”. In business practice, it gives a possibility to build an analytical customer profiles [2], that is a result of mining the time series of customer transactional data, as well as other social-demographic data in order to generate additional business value. In the current approaches, the customer profile is developed and used in marketing campaign management systems. The company possesses a local data warehouse and limits profile creation to customer data analysis accepted by law. In this approach the customer is not informed about data processing that concerns him. In this scenario, a customer analytical profile is a property of the company. The new emerging trend is that selected elements of customer profile start to be traded between companies. The companies exchange information on customer behaviours and customers may be informed about data processing and exchanging and they can use loyalty and benefit programs. In this approach, the analytical customer profile becomes a joint property of companies participating in the program [2]. Such applications of big data analysis have an enormous business potential, but at the same time raise justified questions about protection of personal data.

II. Privacy Issue

The problem of privacy associated with digital footprint registration has in fact two dimensions. The first one is of business character and is associated with the use of customer data to adjust product and/or service offer to the customers’ needs. Such a situation raises the so-called Privacy Paradox [3]. There is a natural contradiction between customers’ will to protect their personal data and their will to receive personalised marketing messages. The two conditions cannot be met at the same time. Hence, paradoxically, a customer may agree upon transferring their private data and even contribute to the process of profile building, in exchange for special offers, bonuses, discounts, promotions, etc. Another dimension of the privacy issue is associated with the customer’s personal life, when companies gain access to particularly sensitive data that go beyond standard customer behaviours and concern e.g. the moral sphere. In this case, the paradox arises from the fact that every individual has a right to not disclose their personal data, but at the same time there is virtually no possibility to live in the modern world without using mobile phones, credit cards or the Internet. The issue seems of even greater importance, given the fact that the registered data may be stored without any time restrictions. Thus, our past actions may impact the present, but also distant future, affecting even our great-grandchildren [4].

It may be stated that the notion of privacy has a great importance for our life and business model development in the era of digital economy [5] [6]. It was decided to examine this important phenomenon based on the behavioural data gathered via Facebook, a popular on-line social network site. Fundamental features to the experience on Facebook are a person's newsfeed which is a personalized feed of his or her friends updates. Facebook requires the user to identify themselves authentically. The profile displays information about the individual he or she has chosen to share, including interests, education and work background and contact information. Facebook also includes core applications – photos, events, videos, groups, and pages – that let people connect and share in rich and engaging ways. Users can communicate with one another through chat, personal messages, wall posts, pokes, or status updates. Additionally Facebook users have a control over the profile visibility and on-line activity.

Related work on the privacy issue in Facebook is widely presented in Stutzman et al. [7] and Ibrahim et al. [8] papers. The research group consisted of college students. The choice of this particular research group was not accidental, since potentially, young people display less interest in personal data protection on a social networking site. The main goal of the research was to examine to what extent the users employ privacy settings and the influence of such actions on their behaviours.

III. RESEARCH SETTINGS AND DATA COLLECTION

Kevin Lewis presented a use of a longitudinal dataset combined with network modeling to examine the co-evolution of college students friendships and privacy behavior on Facebook [9]. He found significant degree effects on students privacy behavior, where students with larger networks are more likely to have a private profile. Private here mean "separate "from strangers. In this context we can formulate the following hypothesis:

"The increase in the number of friends will increase the likelihood of posting status updates where the users are using privacy settings"

We decided to carry out a two-stage empirical study. The first stage is to determine to which extent the users of social network sites deliberately use privacy controls. This stage will be carried out with the use of descriptive statistics. Second stage, on the other hand, is to determine the influence that the usage of privacy controls have on the users' activity. This study is corresponding to the Piskorski social strategy theory [10].

In spring 2011 the Warsaw School of Economics students were invited to participate in research experiment. No course credit was and students were not compensated for the study. We did, however, offer a free participation in the workshop on social media in business applications. We created the project webpage which described the project in complete detail including the Graph API protocol description, which is officially supported by Facebook. The Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). Every object in the social graph has a unique identifier. After two weeks advertising campaign (e-mails, newsletter, students journals, etc.) at Warsaw School of Economics we received a positive response from 272 Facebook users. The received data were anonymized and the whole study was organized in the way to avoid the selection bias in study participation. Our research is based on the 214 active users having the following characteristics (those data are based on the survey answers): age (18-26 years old) – 71%, sex (men) - 51 %, education (college student) – 75%, location (city more than 1 million population) – 72%. We started to download all available data provided by Graph API excluding chat conversation data since June 2011. The data presented in this study were collected in August 2011, two months before the date, when Facebook lunched his new privacy policy introducing pre-defined user groups.

In order to conduct the regression analysis the three kinds of variables were established from data provided by Graph API:

- Dependent variable – user has an opportunity to create content (update status, post a picture, etc.), that is posted to a user Wall (timeline). Then Facebook automatically is generating a News Feed story so that the next time user friends log into Facebook, they see the content the specific uses shared on their homepage. It is possibility control the visibility of these stories by using the audience selector. By means of a dropdown menu it is possible to choose who an user want to share a post with. To capture user activity we constructed the dependent variable *Self posts*. This variable is defined by a number of posts (the possible post types are: photo, status, link, video, check in, music, flash object) that a specific user published on his Wall during an experiment period, and are available to read and comment by his friends.
- Independent variable - we took into account the size of the social network that is straightforward measured by the *Friends number* variable. This variable is measure by a total number of items on the friends list.
- Control variables - we included two types of control variables: user demographics controls and a measure of users interactions. First, we controlled for user demographics data by including a binary variable *Sex*. Second we controlled for the interactions to a specific users from his friends by *Others posts* variable. This variable is measure by a number of posts published on an user Wall by his friends.

Unfortunately, due to Graph API restriction, it was impossible to download the whole friends network of our experiment participants. Due to this restriction we were not able to calculate more advanced social network analysis measures like density. Anyway the results presented in the next chapter are promising.

IV. RESULTS

Almost 75% (160 on 214) of the participants were conscious about privacy setting and they created they own private friends lists. Those results are quite similar to the boyd and Hargittai survey research on Facebook privacy setting [11]. Table I reports the descriptive statistics and correlations between variables. The average number of friends is 263. The largest correlation coefficient between two independent variables is .205 (between *Friends number* and *Other posts*. The reason that those two variables are relatively highly correlated is that in bigger social networks there is higher probability of receiving a post from others. It is worth mentioning the significant (at the .01 level) correlation between the *Sex* and *Other Posts* variables, that interpretation is that women are receiving posts their Wall from others in opposite to men. There is a very intriguing negative correlation (significant on the .05 level) between *Self Posts* and *Other Posts* variables. It means that an intensive activity of the specific subject is not influencing directly others activity on the subject Wall. It should be underline clearly the *Other Posts* variable is not representing other comments to posts.

TABLE I. DESCRIPTIVE STATISTICS AND CORRELATIONS

Variables	Mean	S.D.	1	2	3	4
1. Self posts	54.201	79.910	1			
2. Sex (male=1)	.0523	.0501	.001	1		
3. Others posts	8.893	7.820	-.138*	-.196**	1	
4. Friends number	263.065	179.799	.170*	-.065	.205**	1

Total number of observations = 214; *p < .05 **p < .01 in a two-tailed test

TABLE II. REGRESSION ANALYSIS FOR *SELF POSTS* VARIABLE

Model	1	2	3
Sex (male=1)	-7.38	-13.13	-9.84
	(15.63)	(13.73)	(10.76)
Others posts	1.61	-.82	-51
	(1.27)	(1.02)	(.82)
Friends number	.08	.13*	.11*
	(.06)	(.04)	(.03)
Number of observations	54	160	214
R	.55	.41	.41
R2	.31	.16	.17
F statistic	4.24	5.01	6.95

Numbers in parentheses are standard errors. Constant was omitted. *p < .001 in a two-tailed test

The main goal of this empirical study is to determine the impact of using privacy controls have on the users' posting activity. We were able to establish a user privacy awareness by monitoring usage of the Facebook privacy settings. As it was mentioned by means of a dropdown menu it is possible to choose who an user want to share a post with: Public, Friends of Friends, Friends, and Custom (includes specific groups, friend lists or people an user specified to include or exclude). Table II contains the results from regression analysis in which the dependent variable is *Self posts*. We tested the dependency between user activity (*Self posts)* and the grow of the social network (*Friends number)* in the context of privacy concern in the three models. The first model (column 1 in the table II) was established for the 54 non privacy aware users. The second model (column 2 in the table II) was created for the 160 users which were concern about privacy proving that by using privacy settings. The third model (column 3 in the table II) was established for the whole group. In contrary to the first model, the second one is a significant. The coefficient of *Friends number* is positive and highly statistically significant at .001 level. This exactly means that increase in the number of friends will increase the likelihood of posting status updates where the users are using privacy settings.

V. CONCLUSIONS

The conducted empirical studies are of preliminary character, but give an insight into the phenomenon. The participants of the study showed a relatively high awareness of the use of privacy controls. The research proved that conscious use of privacy controls is influenced by the level of users' activity, which in this case stays in positive relationship with the number of friends on the website. Contrarily to common belief, even group of students, exercise their right to privacy whenever it is possible. Based on the obtained results, it may be supposed that the impossibility to control privacy settings could reduce or even terminate users activity. These conclusions find confirmation in the policy of Facebook, which in Fall 2011 introduced pre-defined users and groups.

As it has been shown, information privacy is a particularly important issue that may be subject to conscious management. In a broader perspective, the phenomenon has much greater implications for all the business use of big data analysis. It may be expected that in the near future, customers will demand not only to control their personal data gathered by the third party, but also to be given access to control mechanisms. Moreover, they might also demand the right to delete given information. This means that ethical development of business models based on a wide repertoire of customers' behavioural data cannot proceed without taking into consideration the active role played by customers and their consciousness of such revolutionary changes.

ACKNOWLEDGMENT

This research was financed by the National Science Centre in Poland under grant number UMO-2011/01/B/HS4/06630. The author would like to thank Mikolaj Jan Piskorski for his valued comments on the privacy issue in social on-line networks and Mateusz Chrzest for his help in data collection. Additionally the author would like to thank the anonymous reviewers for their helpful and constructive comments.

REFERENCES

[1] A. Petland, Honest Signals: How They Shape Our World. Cambbridge: MIT Press, 2010.

[2] J.Surma, “Modeling customer behavior and social relations with analytical profiles” in Social Network Mining, Analysis and Research Trends: Techniques and Applications, I-Hsien Ting, Eds. New York: IGI Global, 2012, pp. 171-182.

[3] S. Kelly, Customer Intelligence. From Data to Dialoque. New York: Willey, 2006.

[4] V. Mayer-Schönberger, Delete: The Virtue of Forgetting in the Digital Age. New Haven: Princeton University Press, 2009.

[5] A. Goldfarb, C. Tucker, “Privacy Regulation and Online Advertising”, Management Science, 57(1), 2011, pp. 57-71.

[6] C. Tucker, “Social Networks, Personalized Advertising, and Privacy Controls”, NET Institute Working Paper No. 10-07; MIT Sloan Research Paper No. 4851-10, 2011.

[7] F. Stutzman, R. Gross, A. Acquisti, “Silent Listeners: The Evolution of Privacy and Disclosure on Facebook”, Journal of Privacy and Confidentiality 4 (2), 2012, pp. 7-41.

[8] S. Z. Ibrahim, A. Blandford, N. Bianchi-Berthouze, "Privacy Settings on Facebook: Their Roles and Importance", IEEE International Conference on Green Computing and Communications, 2012, pp.426-433.

[9] K. Lewis, J. Kaufman, N.Christakis, “The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network”, Journal of Computer-Mediated Communication, 14, 2008, pp.79–100.

[10] M.J. Piskorski, Competing with Social Networks: How to Leverage Social Media for Profit. New Haven: Princeton University Press, 2013, in press.

[11] D.M. boyd, E. Hargittai, “Facebook Privacy Settings: Who Cares?”, First Monday, 15 (8), 2010.