Saving Private Data: The Economics Of Online Privacy

Greg Mangan
Senior Sophister

Economic activity is increasingly taking place in an online setting and the security and privacy of people's personal information is increasingly called into question by whistleblowers and media outlets. In this econometric investigation, Greg Mangan uses a logistic regression model to investigate what influences people's concern about privacy. While his results do not support hypotheses about demographic differences he finds that being the victim of a cybercrime has a significant effect on fear. These results suggest that stereotypes about different demographics' attitudes to the internet may be an oversimplification, and that in reality differences in attitude are the result of different experiences.

Introduction

As internet users, should we be worried about our privacy online? Furthermore, are we worried about our online privacy? There are many causes for fear in offering our personal information online. However, from the point of view of businesses, and those concerned with the function of economic markets, it is the fear itself that should be feared.

While much of the fear is justified due to credible threats such as scams and identity theft, there is a significant portion of worry that is unfounded, which impedes the full efficiency of online markets. Understanding the motivation for this fear and the demographic breakdown could be a huge advantage to businesses in terms of how they market their products and how they invest in security measures.

This paper aims to answer the question: *who* is worried about online privacy? Beginning with a literary review of writings on the economics of privacy, a logit model is then introduced to answer this question based on a comprehensive dataset that ties attitudes and experiences of online privacy with demographic information. Through the course of the paper, 'anonymity' will refer specifically to online anonymity unless stated and will be used interchangeably with 'online privacy'.

Literary Review

A very important distinction to make in relation to online privacy is between attitudes

and behaviours. Acquisiti (2004) take a psychological approach to explaining inconsistencies in consumer behaviour, claiming that consumers cannot be expected to act rationally in the decision making process for e-commerce due to self-control problems and the preference for instant gratification. Some users may fail to behave in a way that protects their privacy, even if this conflicts with their attitude towards privacy.

Nehf (2007) furthers the debate over dichotomy between consumer actions and consumer attitudes. A case of self-imposed information asymmetry is proposed in that most web users simply fail to read online privacy policies for websites. These are often lengthy blocks of convoluted prose that effectively mask the important implications for user privacy. Also, individuals may assume that a website suitably protects their personal information from the belief that the brand name is trustworthy, due to their practices and reputations in other (not necessarily online) markets. While companies may use this signalling to their advantage, there is often no basis for this assumption on the individual's part. As such, they are in a position where they have an incentive to accumulate personal information (which can generally be used to analyse preferences and inform profit-maximising policies) with little backlash from their users, who are therefore presented with a genuine cause for concern over their information availability online.

Miyazaki and Fernandez (2001) analysed studies in the area of risk perceptions and how they affected individual's involvement in online markets from the perspective of 'internet literacy'. They found evidence in support of their hypothesis that 'internet experience is positively related to the rate of purchasing products online'.

While freedom and choice are properties that are highly valued by economists, it is argued in 'Privacy and Freedom: An Economic (Re-)Evaluation of Privacy' (van Aaken, Ostermaier and Picot, 2014) that privacy (which hasn't always been given the same weight of importance) is a type of freedom and should therefore be considered a fundamental economic concept. The authors construct an argument built around economic liberalism and the idea that 'freedom has intrinsic value'. Revocability is identified as a key requirement in terms of individuals giving up their privacy- one should be able to reclaim it. Many websites that store large amounts of personal information such as Facebook and Google fail to offer or weakly uphold this concept of revocability; the paper commends the EU's attempts to introduce a 'right to be forgotten'.

Setting aside the question of whether individuals have cause for concern, the mere fact that they have concern is detrimental to the functioning of eCommerce markets. A Federal Trade Commission report given to the US congress (FTC, 2000) noted that studies show 'privacy concerns may have resulted in as much as $2.8 billion in lost online retail sales in 1999, while another suggests potential losses of up to $18 billion by 2002'. Unjustified worry may be considered a market imperfection, a failure to 'sustain desirable activity' (Bator, 1958). 'Desirable activity' in this context is the highly eff-

icient online market for goods. Understanding worry over privacy issues, whether justified or unjustified, is therefore essential for all firms engaged in this economic activity.

Empirical Approach

Data

To conduct an empirical investigation into the determinants of concerns over online personal information, a cross-sectional data set from the Pew Research Institute was selected. The data was collected from a survey on anonymity, privacy and security online (Pew Research Institute, 2013). The majority of the questions were only asked of respondents who initially answered yes to either being an internet user or a smartphone user, and so the empirical analysis has been restricted to this subset of individuals. Most of the variables are binary dummy variables, derived from questions with 'yes' or 'no' answers.

The important variable that we seek to explain is that of whether an individual is worried about their information being online or not. Standard personal information such as age and sex is included for each respondent. Respondents are also asked if they are a parent of a child 18 or younger. Other explanatory variables can be gathered from the survey such as views on the possibility of anonymity, views on the right to anonymity and attempts at anonymity. The final data we will draw from the survey comes from a set of questions about negative online experiences, for which we will consider a new binary dummy variable that represents whether or not an individual has been a victim of one or more of the listed abuses.

Model

The variables to be used in the model are:

Yi = WORRIED-A binary dummy variable that takes the value of 1 if the respondent answers yes to *'Do you ever worry about how much information is available about you on the internet…?'*

X1 = PARENT-A binary dummy variable that takes the value of 1 if the respondent is a parent or guardian to a child under 18

X2 = AGE-The age of the respondent in years

X3 = SEX-A binary dummy variable that takes the value of 1 if the respondent is male and 0 if female

X4 = ANON_RIGHT-A binary dummy variable that takes the value of 1 if the respondent answers yes to *'Do you think that people should have the ability to use the internet completely anonymously for certain kinds of online activities?'*

X5 = ANON_POS-A binary dummy variable that takes the value of 1 if the respondent answers yes to *'…do you think it is possible for someone to use the internet completely anonymously…?'*

X6 = ANON_TRIED-A binary dummy variable that takes the value of 1 if the respondent answers yes to *'Have you ever tried to use the internet in a way that hides or masks your identity…?'*

X7 = VICTIM-A binary dummy variable that takes the value of 1 if the respondent answers yes to at least one of eight questions about being the victim of a mishap due to online activity (stolen data, account compromisation, scam, harassment, loss of job/education opportunity, relationship trouble, reputation damage and physical danger)

To estimate the effects of the variables on individuals' worries we define the following logistic model:

$$P(Y_i = 1) = exp(Z_i)/[1 + exp(Z_i)] \qquad for\ Zi = \sum_{k=0}^{j} \beta_k X_{k,i} \ (X_{0,i}=1 \forall i)$$

The model was run firstly for j=6 and secondly for j=7.

Choice of Statistical Model

The most important point to note from the variable specification above is that the dependent variable is binary. Therefore, the aim is to build a model that predicts P(Yi=1), the probability of a yes for WORRIED. The linear probability model is one method for achieving this, which follows a standard OLS regression approach. However this gives rise to two notable issues: the predicted values may fall outside of the range of the closed set [0,1] (meaning an undefined probabilistic interpretation) and the marginal effect of changes in explanatory variables is assumed to be constant (Wooldridge, 2012). Instead, the logit model is used as an alternative.

The logit model is built around a cumulative distribution function (CDF), as this necessarily maps onto [0,1], solving the first issue above. The probit model is also built around a CDF of the normal distribution. The probit would give similar results, especially given the large sample size (n=770). Seeing as the CDF used is a function of exponentials,

the marginal effects are no longer simple constants, as was the case with linear models. For the logit model 'the magnitude of the effect varies with the values of the exogenous variables' (Aldrich and Nelson, 1984), which solves the second issue above of the unrealistic assumption of a constant effect.

Expectations

Firstly, the data may be summarised as below:

Variable	Obs	Mean	Std. Dev.	Min	Max
WORRIED	770	.4844156	.5000819	0	1
PARENT	770	.2779221	.4482659	0	1
AGE	770	48.57143	17.52073	18	93
SEX	770	.5064935	.5002828	0	1
ANON_RIGHT	770	.5909091	.4919857	0	1
ANON_POS	770	.3506494	.4774835	0	1
ANON_TRIED	770	.1584416	.3653919	0	1
VICTIM	770	.3584416	.4798544	0	1

Table 1: Data Summary

Two groups that could be assumed to be sceptical of online safety would be parents and the elderly. Parents may be more actively concerned, as concern for their child's online safety may force them to more strongly consider the dangerous of their own information being online. Due to the rapid pace of technological change, the elderly are more likely to be unfamiliar with new technologies and therefore possibly more sceptical of them. Therefore, it is expected that the results will show-positive relationships with WORRIED for both PARENT and AGE. SEX is included as a control variable, aiming to reduce omitted variable bias. It is not expected that it would have any important effect on WORRIED.

It is expected that an individual who has tried to mask their online activity at some point would naturally be more likely to have concerns about online anonymity. As such, a significant positive impact on WORRIED is expected of the ANON_TRIED variable. The two other dummy variables related to questions posed about anonymity are less clear-cut. The effects of the variables ANON_RIGHT and ANON_POS are therefore ambiguous.

A stereotype of an individual who is worried about their information being available online may have historically leaned towards an image of paranoia. However, given the rise of cybercrime in recent years (RTÉ, 2015) there is much more justifiable cause for concern. As such, it is expected that VICTIM will have a significant positive impact on

WORRIED.

Results

Logit Interpretations

Interpretation of coefficients in the logit model is different to that of standard OLS regression due to the non-linear relationship between dependent and independent variables. The interpretation is 'less straightforward' (Aldrich and Nelson, 1984). However we can still say that the sign of the coefficient determines the direction of the effect and that greater magnitudes correspond to larger effects. The main difference is that we cannot state the effect on the dependent variable of a per unit change in an explanatory variable, we can only give statements such as: 'an individual who answered yes for Xi, is more/less likely to be worried'. Significance is discussed in relation to the 5 per cent level.

Interpreting the Results

The logit test was run first for the case of j=6 (i.e. just using the first six explanatory variables listed in the model section), giving the following output:

```
Logistic regression                               Number of obs   =        770
                                                  LR chi2(6)      =      21.12
                                                  Prob > chi2     =     0.0017
Log likelihood = -522.78899                       Pseudo R2       =     0.0198
```

WORRIED	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
PARENT	.0709351	.1715258	0.41	0.679	-.2652493	.4071196
AGE	-.0010371	.0045066	-0.23	0.818	-.0098699	.0077957
SEX	-.2610505	.1493497	-1.75	0.080	-.5537705	.0316694
ANON_RIGHT	.2606078	.1539064	1.69	0.090	-.0410432	.5622588
ANON_POS	-.2474577	.1549434	-1.60	0.110	-.5511413	.0562259
ANON_TRIED	.6701472	.2083145	3.22	0.001	.2618583	1.078436
_cons	-.0720176	.2962919	-0.24	0.808	-.6527392	.5087039

Table 2: Logistic Output for j=6

The first point to note is that the likelihood ratio chi-squared and its associated p-value of 0.0017 mean that the model is significantly better than a model with no predictors. As expected the ANON_TRIED variable has a strong positive impact. The p-value of 0.001 suggests that it is highly significant. This would seem to go against the idea of Acquisiti (2014) that there exists a dichotomy between beliefs and behaviour as regards online anonymity.

ANON_RIGHT is not statistically significant for a one-tailed test, but using a one-tailed test gives a p-value of 0.090/2=.045, which is statistically significant. An ar

gument for this would be that it would seem illogical for an individual who does not believe in a right to anonymity to then be worried about their personal information being online. ANON_POS does not appear to be significant.

Surprisingly, PARENT and AGE seem to have very little statistical significance, which would challenge the perception of parental and elderly scepticism of technology. Even more surprising is that SEX, while not statistically significant, is not too far off with a p-value of 0.08. If significant, the coefficient would suggest that females tend to be more worried. It would be interesting to see if a test on a larger sample would give a statistically significant result.

Secondly the model was run for j=7 (the same model as above but this time including the VICTIM term), giving the following output:

```
Logistic regression                             Number of obs   =        770
                                                LR chi2(7)      =      27.03
                                                Prob > chi2     =     0.0003
Log likelihood = -519.83417                     Pseudo R2       =     0.0253
```

WORRIED	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
PARENT	.0516411	.1724416	0.30	0.765	-.2863382	.3896204
AGE	.0010332	.004605	0.22	0.822	-.0079925	.0100589
SEX	-.2558089	.1499272	-1.71	0.088	-.5496608	.0380429
ANON_RIGHT	.272266	.1545852	1.76	0.078	-.0307155	.5752474
ANON_POS	-.2724322	.155989	-1.75	0.081	-.578165	.0333006
ANON_TRIED	.6440991	.2092815	3.08	0.002	.233915	1.054283
VICTIM	.3822288	.1575726	2.43	0.015	.0733921	.6910655
_cons	-.3011752	.3122359	-0.96	0.335	-.9131463	.3107958

Table 3: Logit Model for j=7

Similar interpretations are possible given this test. PARENT and AGE are not statistically significant. AGE has a higher p-value of 0.088 this time, suggesting that in the previous model it was capturing some of the explanatory power of the VICTIM variable. Our pseudo R-squared of 2.53 per cent suggests that it is a relatively small proportion of differences that is being explained by the model; nonetheless some of the variables are statistically significant.

The main outcome from this test is that VICTIM is quite statistically significant factor, due to the p-value of 0.015. The positive effect of this binary variable is as expected, demonstrating that individuals who have been victims of online wrong-doings are more likely to be worried. Paranoia could be interpreted as VICTIM not having a significant impact on WORRIED, and so there does not appear to be an indication of this.

Issues with the Model

One issue with the above question of paranoia is the direction of causality. The most logical statement for paranoia would be a low P(VICTIM | WORRIED), as this looks at the probability that an individual has been a victim given that there are worried. The problem is that our model is structured the other way around and so while our interpretation above could possibly be true, in that paranoia could be indicated by a low significant impact of VICTIM on WORRIED, it is probably safer to say that the model is inconclusive about the existence or paranoia (rather than confirming its absence).

The strongest issue to highlight is in relation to the survey answers being interpreted as 'Yes' or 'No' values. In determining values for WORRIED, ANON_RIGHT, ANON_POS, ANON_TRIED and VICTIM, the 'No' interpretation (the zero binary value) also incorporates those respondents that answered 'Don't know' or those that refused to answer. Therefore the most explicit interpretation would not actually be 'Yes' and 'No', but rather 'Responded Yes' and 'Did not respond Yes'. The two human factors that may distort the results in this case are honesty and willingness to disclose information. In particular the VICTIM variable may be prone to bias in that respondents who were in fact victims of the listed online attacks may, for reasons of shame or denial, answer 'No' or refuse to answer, which would give rise to a significant bias.

Another issue is that our dependent variable is binary, though worry may not be a binary concept. We are assuming that individuals are either worried or not worried. It may be that there are different degrees of worry, even different types of worry, that should be investigated.

Possible Extensions

With the worry issue in mind, one possible extension would be to gather categorical data for worry. To look at the issue from the attitude versus actions perspective of Acquisiti (2014) and the FTC (2002) report, the WORRIED variable could take on distinct values for responses such as 'Yes I am worried but No it doesn't negatively affect my eCommerce activity' and 'Yes I am worried and Yes it does negatively affect my eCommerce activity'.

To further analyse the interaction of the variables in this model, it would be interesting to incorporate questions regarding actual eCommerce engagement. A measure of spending on online purchases could be introduced, for example an individual's estimation of their annual spending in such markets. This could help to paint a clearer image of the type of consumer that is worried about online privacy. Further questions could be included about internet literacy, which Miyazaki and Fernandez (2001) would seem to suggest is an important factor.

A final extension would be a panel data analysis, introducing a time series by repeated surveying of the fixed sample of individuals. It would be very interesting to analyse

the change in worries over time, and to single out the effects of large scale privacy scares such as the Edward Snowden NSA whistleblowing incident (Brown, 2014).

Conclusion

Though the paper has not managed to identify key demographics that are most prone to worry over piracy, this in itself could be a lesson to businesses. There is no singular face of online fear, and so efforts to address privacy concerns should not depend on the demographic of a firm's customer base.

The strongest result from the empirical analysis confirms that victims of online attacks are more likely to be worried about their privacy, and also that there seems to be evidence that individuals attitudes regarding privacy concerns are in fact aligned with their behaviours.

Finally, issues that could cause bias in the results are raised, giving rise to a discussion of an extended survey that would incorporate factors such as a non-binary view of worry and an inter-temporal element, which would potentially give clearer results to the question of who is worried about online privacy.

References

Acquisti, A. 2004. Privacy in electronic commerce and the economics of immediate gratification. In Proceedings of the 5th ACM conference on Electronic commerce, pp. 21-29. ACM.

Aldrich, J.H. and Nelson, F.D. 1984. Linear probability, logit, and probit models (Vol. 45). Newbury Park: Sage University Papers.

Bator, F.M. 1958. The anatomy of market failure. The Quarterly Journal of Economics, 72:3:351-379.

Brown, M. 2014. Edward Snowden: the true story behind his NSA leaks. The Guardian. [on-line], http://www.telegraph.co.uk/culture/film/11185627/Edward-Snowden-the-true-story-behind-his-NSA-leaks.html [Accessed: 23 October 2015].

FTC. 2000. Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. [on-line], https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000text.pdf [Accessed: 28 October 2015].

Miyazaki, A.D., and Fernandez, A. 2001. Consumer perceptions of privacy and security risks for online shopping. Journal of Consumer affairs, 35:1:27-44.

Nehf, J. P. 2007. Shopping for Privacy on the Internet. Journal of Consumer Affairs, 41(2), pp. 351-375.

Pew Research Institute. 2013. 'July 2013–Anonymity(Omnibus)'. [on-line], http://www.pewinternet.org/datasets/july-2013-anonymity-omnibus/ [Accessed: 5 October 2015].

RTÉ. 2015. Survey shows cyber crime on the rise. [on-line], http://www.rte.ie/news/2015/0211/679303-cyber/ [Accessed: 28 October 2015].

Van Aaken, D., Ostermaier, A., and Picot, A. 2014. Privacy and Freedom: An Economic (Re−) Evaluation of Privacy. Kyklos, 67:2:133-155.

Wooldridge, J.M. 2012. Introductory Econometrics: A Modern Approach (5 ed). Ohio: Cengage.