An Exploratory Analysis on the Risk to be Offended on the Internet*

Susanne Kirchner Leopold Sögner†

October 6, 2016

Abstract

Questionnaire data is used to identify socio-demographic as well as the risk-awareness characteristics of users offended on the Internet. The data comprises a representative sample of 3,000 individuals, containing information on employment, education, age, the frequency of Internet usage and security measures taken by the users. By means of a cluster analysis, within the sub-sample of offended users, we identify a female group, where employment and education are high, a male cluster with similar characteristics, a group of urban users with low security awareness and a group of young users. Regressions show that the frequency of using the Internet increases, while to communicate only to people known in real life reduces the risk to be offended on the Internet.

Keywords: Cyber-Crime, Cluster Analysis, Internet Victimization.

*The authors thank B. Angleitner, M. Gstrein, U. Röhsner (MAKAM Research), M. Popolari and A. Mattern (both Austrian Federal Ministry of the Interior; Sektion IV (Department IV/6)) as well as especially Robert Kunst for interesting discussions and comments. We gratefully acknowledge funding from the Austrian KIRAS program (KIRAS security research program) financed by the Austrian Federal Ministry of Transport, Innovation and Technology.

†Susanne Kirchner (kirchner@ihs.ac.at), Tel. + 43 1 59991 185, Leopold Sögner (soegner@ihs.ac.at), Tel. + 43 1 59991 182, Department of Economics and Finance, Institute for Advanced Studies, Josefstädter Straße 39, 1080 Vienna, Austria. Leopold Sögner has a further affiliation with the Vienna Graduate School of Finance (VGSF).

1 Introduction

The growing number of providers and users of Internet services as well as social communication networks also raises security issues and resulted in the emergence of cyber-crime research (see, e.g., Hartel et al., 2011). This article uses questionnaire data and applies statistical methods to identify users and their characteristics who were subject to some form of offense on the Internet and Social media by means of a cluster analysis. In a second step, we analyze how these characteristics are related to the likelihood of being offended on the Internet. In particular, we investigate how different safety measures taken by users affect their protection against cyber-crime.[1]

In parallel to the emergence of cyber-crime science, for governmental institutions, such as the Ministry of the Interior and police authorities, the criminal aspects (such as theft of data, hacking, fraud, etc.) have become of particular interest (see, e.g., the study of Kirchner et al., 2015, instructed by the Austrian Federal Ministry of the Interior). To implement policies with the goal to improve cyber-security and to reduce crime (see, e.g., Becker, 1968; Freeman, 1999; Hartel et al., 2011; Dimkov, 2012), knowledge about the actual number of crimes committed, the socio-demographic structure of the users offended, factors (variables) raising the probability of an offense as well as the cost of Internet crime becomes important (for a cost-benefit for hackers and the cost of cyber-crime see, e.g., Kshetri, 2010; Anderson et al., 2013; Cook et al., 2014).

In addition, also governmental as well as non-governmental institutions provide guidelines how to responsibly use information technology. Such commandments are e.g. provided by CPSR (2015), E.C. (2016) or OeIAT (2016). For these institutions, knowledge on the socio-demographic structure of the users offended as well as user characteristics connected to offenses can be helpful to provide target group specific information, with the goal to increase the risk-awareness and to reduce the risk to be offended. Regarding the effectiveness of security awareness, Bullée et al. (2015) showed in experiments that measures to increase security awareness turned out to be statistically significant.

Let us relate this article to recent literature: An overview on recent developments and results in cyber-crime science is e.g. provided in Hartel et al. (2011) and Dimkov (2012). Regarding academic publications in the field of cyber-crime research, almost recently, Hartel et al. (2011) intensively searched through literature in various academic disciplines and concluded that "In spite of our efforts we have failed to find

[1] For some more detailed definitions on cyber-crime see Appendix A.

documented scientific studies of how Information Security effectively prevents cyber-crime." By looking for causes of this gap, the authors claim that problems in information security are hardly reported to the police for several reasons. For example, a problem in the information security system can but need not result in crime or firms try to solve problems internally.

Cyber-bullying was investigated in the empirical study of Hinduja and Patchin (2008). The authors used an on-line survey tool to collect data from 6,800 users in the time span December 2004 to January 2005. After focusing on the group of users not older than 17 years and data cleaning, the authors ended up with data from 1,378 users. The response variables constructed by the authors are two victimization variables ("general/serious cyber-bullying victimization") and two offending variables ("general/serious cyber-bullying offending"). Regarding serious cyber-bullying victimization, the authors observe (by applying logistic regression) that the time spent at the computer, school problems and being a bullying victim in real life are positively related to victimization. Other variables such as gender, age, black/white and peer effects turned to be insignificant. Due to the different age structure of the users, relating the study of Hinduja and Patchin (2008) to the results obtained in this article is difficult.

Information security awareness of Internet users was analyzed in Tsohou et al. (2008) as well as Talib et al. (2010). While Tsohou et al. (2008) provide an overview on information security awareness, the study of Talib et al. (2010) is based on survey data containing 333 observations. The authors argue that – compared to private use — at an individual's workplace clearer legislation and regulation about IT security exist. Because of this, the authors claim that learning about Internet security mainly takes place at an individual's workplace. Then, positive spill-over effects to security awareness at home are observed. Moreover, information on the "the perception of security in e-commerce B2C (business to customer) and C2C (costumer to customer) websites" is provided by Halaweh and Fidler (2008), who followed a qualitative approach by interviewing fifteen customers and twelve organizations' managers and their IT staff.

Kirchner et al. (2015) analyzed which criminal-relevant phenomena and activities do occur in social media, to what extent did they reach so far, and which methods to attack users were applied. By using questionnaire data, containing information from 3,000 individuals, the study shows that Facebook (used by 62% of the people asked in the questionnaire), WhatsApp (50%) and YouTube (46%) are those social media, which are used most frequently in the age group 14 - 49 years old. Regarding police relevant issues, Kirchner et al. (2015) observed that defective software/malware, hacking, fake accounts, cyber-mobbing (see also Schneider et al., 2013, and the literature cited there), phishing, cyber-bullying (see also Hinduja and Patchin, 2008), cyber-stalking, profile copying, sexting (see also Lee et al., 2013), and happy slapping are the most frequent ways how users were offended (the order of these terms corresponds to their frequency of occurrence).

This article uses the questionnaire data collected by Kirchner et al. (2015) and identifies groups of offended Internet users. In particular, Section 2 describes the data. To obtain information on the security-awareness and the socio-demographic characteristics of the users offended, Section 3 first presents results obtained by means of a cluster analysis. In a second step logit and probit regressions are performed to investigate the impact of user characteristics on the risk to be offended on the Internet. Section 4 concludes.

2 Data

A very first step to investigate the risk of being offended on the Internet is to look on the number of notifications and complaints collected by police authorities. For example, the Austrian Ministry of the Interior collects the number of notifications on a yearly basis (for Austria, see e.g., BM.I, 2015, "Austrian Security Report"). This report shows the following: For 2014 a decline in the area of Internet crime is reported (-10.8% compared to 2013), while for the last decade an increase from 1,794 notified offenses in 2005 to 8,966 notified offenses in 2014 is observed. After the significant rise in the last decade and the decrease in 2014, the criminal offenses are less than 10,000, which corresponds to approximately 0.1% of the total Austrian population. The number of notified offenses is to be found mainly in the area of cyber-crime in a broader sense, and particularly, in the field of Internet fraud.

During the same periods, also the number of complaints increased enormously. In particular, from 1,151 in 2005 to 7,667 complaints in 2013. In parallel to the number of notifications, the complaints with respect to Internet fraud fell by 13.5% in the year 2014. However, the value of 6,635 complaints in 2014, is imperceptibly higher than the value in 2012, where 6,598 complaints were observed. In addition, police authorities are also concerned about a large dark field in the area of cyber-crime, and point out that new criminal phenomena are in progress (see Bundeskriminalamt, 2015).

To obtain more detailed information, this article uses data from the study of Kirchner et al. (2015),

Table 1: Descriptive statistics.

	Total	N.A.	$\mathcal{O}$
Men	1510	6	248
%	50.3%	0.4%	16.4%
Women	1490	16	222
%	49.7%	1.1%	14.9%
Total	3000	22	470
%	100.0%	0.7%	15.7%

[a] Number of male and female participants in the study of Kirchner et al. (2015). Sample size $N = 3,000$. The table presents the number of men and women who were already confronted with cyber-crime. $\mathcal{O}$ stands for personally confronted with cyber-crime, while *N.A.* stands for no answer.

where data on socio-demographic factors as well as on offenses on the Internet were collected for a target group of $N = 3,000$ representative users with an age between 14 and 49 years (more details on the data collection process are provided in Appendix B). Table 1 presents some descriptive statistics obtained from this questionnaire data. For the sample of $N = 3000$, the number of people personally confronted with cyber-crime is $\mathcal{O} = 470$. Comparing the rate $\mathcal{O}/N \approx 16\%$ to the notification rate of approximately 0.13%, based on the data provided in BM.I (2015)[2] , strongly supports the arguments provided e.g. in Appendix A of Hartel et al. (2011), who claimed that the number of offenses is above the number of offenses notified by the police. The differences observed between the male and the female population turned out to be small (this difference is also statistically insignificant at a 5% significance level).

Next, the data collected by Kirchner et al. (2015) is used to construct $k' = 21$ variables. In more formal terms, the data $\mathbf{X}' \in \mathbb{R}^{N \times k'}$ contains the observations $\mathbf{x}'_n = (y_n, x_{n2}, \ldots, x_{nk'})^\top \in \mathbb{R}^{k'}$ for $n = 1, \ldots, N = 3,000$.[3] The variables y_n, x_{ni}, $i = 2, \ldots, k'$, are:

y_n : The binary variable *Attacked*, where 0 implies that the corresponding individual was not personally offended on the Internet or social media, while the variable is 1 if the user was offended personally. Hence, $\mathcal{O} = \sum_{n=1}^{N} y_n$.

x_{n2} : The variable *Frequency*, measuring the frequency of Internet and social network usage. This variable is an integer ranging from 0 to 2. The value 0 stands for no current use of social networks, 1 stands

[2]To obtain an estimate of the notification rate, we use the Austrian population ($\approx$ 8.5 million) in 2014 and an estimate of the percentage of users in Austria (82%) in 2015, which is supposed to be a good approximation for the year 2014, from Statistik Austria (2015). Then $8966/(0.82 \cdot 8500000) \approx 0.00129$.

[3] For vectors and matrices boldface notation is applied. That is to say, $\mathbf{z} \in \mathbb{R}^p$ denotes a p-dimensional column vector, $\mathbf{Z} \in \mathbb{R}^{p \times q}$ a $p \times q$ matrix. z_i stands for the i-th coordinate of the column vector $\mathbf{z}$. $\mathbf{z}^\top$ denotes the transpose of $\mathbf{z}$.

for occasional use and 2 for frequent use.

x_{n3} : The binary variable *Gender*, where 0 stands for male and 1 for female.

x_{n4} : The integer variable *Age*, measured in years.

x_{n5} : The variable *Inhabitants* approximates the number of inhabitants of the city where the individual currently lives. Here, the following categories are used: 1 stands for $< 10,000$ inhabitants, 2 stands for more than or equal to $10,000$ and less than $50,000$ inhabitants, 3 stands for more than or equal to $50,000$ and $< 100,000$ inhabitants, 4 stands for more than or equal to $100,000$ and $< 250,000$ inhabitants, while 5 stands for $\geq 250,000$ inhabitants.

x_{n6} : The integer variable *Employment* denotes the current employment status, where 0 stands for unemployment, 1 for part time employment and 2 for full employment. On leave, retirement, apprenticeship, civil- or military service and pupils are treated as missing values.

x_{n7} : The variable *Human Capital* (Education), measuring the highest level of education obtained by individual n. This variable is equal to 1 if no school was completed, to 2 if the highest degree is from a secondary modern school ("Pflichschulabschluss in the Austrian school system), to 3 if an apprenticeship, a school without general qualification for university entrance ("Berufsbildende mittlere Schule" or "Allgemeinbildende höhere Schule ohne Matura" in the Austrian school system) was completed, to 4 if a grammar school or an equivalent degree ("Berufsbildende höhere Schule" (e.g., HAK, HLW, HTL) in the Austrian school system) was completed, while 5 stands for some university degree (or (almost) equivalent degrees like "Abiturientenlehrgang, Kollege, Pädagogische Akademie" in the Austrian education system).

$x_{nS,j}$: Binary *Security/Incertitude* variables: The variable $x_{nS,j}$, $j = 1, \ldots, 14$, is set to 0 if an individual did not consider the corresponding security issues as relevant, while the value of the variable is one if the individual cared about that particular Internet security issue. The variables considered are $x_{nS,1}$, 'adapt protection settings at the first registration', $x_{nS,2}$, 'regularly change password', $x_{nS,3}$, 'use different passwords at various platforms', $x_{nS,4}$, 'install safety software', $x_{nS,5}$, 'do not use unsecured WLAN connections', $x_{nS,6}$, 'only communicate with persons known in real life', $x_{nS,7}$, 'never provide personal information', $x_{nS,8}$, 'read terms and conditions carefully at every registration', $x_{nS,9}$, 'deactivate automatic save password facilities', $x_{nS,10}$, 'delete cookies', $x_{nS,11}$, 'hide/tape microphone and camera', the incertitude variable $x_{nS,12}$, 'use common sense', the security variable $x_{nS,13}$, 'do not use social networks', as well as the incertitude variable $x_{nS,14}$, 'user does not care about any security issues'.

Sample means and standard deviations for the variables y_n and x_{ni}, $i = 1, \dots, 7$, are provided in the last column of Table 2, while the sample means and standard deviations as well as correlation coefficients of $x_{nS,j}$, $j = 1, \dots, 14$, are provided in Table 6 in the Appendix B. If no answer is provided or if the answer "don't know" is chosen for some variable by individual n, we obtain a missing value. For y_n, x_{n2} and x_{n3} no missing values are observed. For the variables age, inhabitants and human capital two, thirty and eighteen missing values are observed. For the variable *Employment* where on leave, retirement, apprenticeship, civil- or military service and pupils are treated as missing values we get 616 missing values, while for each of the security/incertitude variables $x_{nS,j}$ 681 missing values are observed. For the cluster analysis performed in Section 3 all $N = 3,000$ observations can be used by setting the contribution for the corresponding variable to zero when obtaining the distance function, while for the regression analysis observations $\mathbf{x}'_n$ containing missing values were excluded by the software package.

3 Results

This section investigates the questions: (i) 'What groups of persons show an insufficient problem-consciousness concerning cyber-crime and thus being at particular risk?' and (ii) 'What variables increase/decrease the risk to be offended on the Internet?'. Regarding the first question we perform a cluster analysis, while the second question is investigated by means of regressions.

Hence, the first goal of this exploratory analysis is to group (cluster) the data described in Section 2, such that the individuals in the same cluster have stronger similarities than the individuals collected in the other clusters. To perform the cluster analysis in a more parsimonious setting and to avoid similarities in $x_{nS,1}, \dots, x_{nS,14}$ to dominate the clustering results, the security variable $x_{nS,1}$ is selected from $x_{nS,1}, \dots, x_{nS,14}$ when performing the cluster analysis. Hence, the observations used to perform the cluster analysis are $\mathbf{x}_n = (y_n, x_{n2}, \dots, x_{n7}, x_{nS,1})^\top \in \mathbb{R}^k$, where $k = 8$, for $n = 1, \dots, N = 3000$. The data used to perform the cluster analysis is abbreviated by $\mathbf{X} \in \mathbb{R}^{N\times k}$, collecting the observations $\mathbf{x}_n$, $n = 1, \ldots, N$. Additionally, a distance function measuring the dissimilarity between the observations $\mathbf{x}_n$ and $\mathbf{x}_m$ has to be chosen. In this section we apply l_1-distances (= sum of absolute distances or Manhattan distances; see equation (3) in the Appendix C). To measure dissimilarities between clusters the "unweighted pair-group average method" is used (see equation (5) in the Appendix C).

In this article we apply *agglomerate hierarchical clustering* techniques, which start with N clusters (i.e. each observation n is a cluster) and then, based on the distance between groups, groups are merged. This merging procedure is continued until one cluster (containing all elements of $\mathbf{X}$) is remaining. In particular, the agglomerate hierarchical clustering algorithm `agnes` described in Kaufman and Rousseeuw (1990)[Chapter 5] and implemented in the software package `R` by Maechler et al. (2015) is applied. For more details see Appendix C.

By applying this clustering technique to our data $\mathbf{X}$, we observe a high agglomerative coefficient of $AC = 0.98$ (see equation (6) in Appendix C), measuring the quality of the clustering method applied to the data. Based on the dendrogram (see Figure 1 in Appendix C) and with the goal to get a parsimonious description of the data, we decided to present the result where the data $\mathbf{X}$ is clustered into twelve groups. This decision is based on the observation that for the branches on the top of the clustering tree larger differences are observed, while for a larger number of clusters the differences in the variables of interest for this study become small.[4]

Table 2 presents results when $I = 12$ groups are considered. The columns 2 to 13 present the group-specific mean values and the group-specific standard deviations within the corresponding cluster $\mathscr{C}_i$. The last column presents the sample means and the sample standard deviations for each variable, obtained from $N = 3,000$ observations. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$, $i = 1, \ldots, I = 12$. Note that the mean value for the variable *Attack* corresponds to the percentage of the individuals offended on the net, i.e. $\frac{\mathcal{O}}{N} = \frac{470}{3000} = 0.1567$.

In the following we focus on individuals who have been offended on the Internet or social media. From Table 2 we observe that all individuals offended on the Internet are contained in the clusters $\mathscr{C}_6$, $\mathscr{C}_7$, $\mathscr{C}_9$, $\mathscr{C}_{10}$ and $\mathscr{C}_{11}$. These clusters only contain offended users (note that the within-group sample standard deviations of the variable *Attacked* are zero). By adding up these numbers we get 470.

[4]Results for $I = 4, 8$ and 16 groups with l_1 distances and for $I = 4, 8, 12$ and 16 groups with Euclidean distances are provided in Appendix D.

Regarding the socio-demographic factors as well as risk-awareness we observe the following: Class $\mathscr{C}_6$ contains almost only women (the mean of the group-specific gender variable is 0.923), who have a mean age around 34 years and within group standard deviation for the variable age of 9.282, i.e. the age structure of this class approximately corresponds to the age structure of the full sample. In addition, the members of $\mathscr{C}_6$ live in smaller cities and have in the mean a high level of education as well as employment. The class specific mean of the variable *Security* $x_{nS,1}$ is close to the mean of the full sample (see last column).

The majority in class $\mathscr{C}_7$ is male. The group-specific means of the variables *Age*, *Inhabitants* and $x_{nS,1}$ are close to the values in cluster $\mathscr{C}_6$. The group-specific means of the variables *Employment* and *Human Capital* are slightly smaller than the values in group $\mathscr{C}_6$.

Class $\mathscr{C}_9$ contains users who have been offended and live in larger cities, in particular, mainly in Vienna. For cluster $\mathscr{C}_9$, with the exception of the size of the city, most group-specific means are almost the same as the group-specific means of the full sample (given the standard deviations of these variables), however for the users in cluster $\mathscr{C}_9$ the security awareness measured by the variable $x_{nS,1}$ is very low.

Class $\mathscr{C}_{10}$ contains young users who were offended. Last but not least, Class $\mathscr{C}_{11}$ contains only four group members. For these users we get the contradicting result that these users were offended ($y_n = 1$) although they did not use the Internet ($x_{n2} = 0$). In addition, the security awareness of these persons is high ($x_{nS,1} = 1$). This contradicting result can either be explained by mis-reporting (e.g. some interviewees hardly using the Internet reported that they currently do not use the Internet) or that these users changed their behavior *after* they have been offended.

After we have identified classes of offended users and their characteristics, we investigate the second question on variables increasing or decreasing the risk to be offended on the Internet. Given our data set we analyze how the variable *Attacked*, i.e. y_n, is affected by the variables *Frequency*, *Gender*, *Age*, *Inhabitants*, *Employment*, *Human Capital* and the *Security*/*Incertitude* variables $x_{nS,j}$. For example, this allows to investigate the questions whether and how the probability to be offended on the Internet is affected by gender, by age, the security awareness, etc.

To investigate these questions we have to account for the fact that y_n is a binary variable. In formal terms we consider the events $\{y_n = 1\}$ and $\{y_n = 0\}$. Logit and probit regressions (see, e.g., Greene, 1997; Cameron and Trivedi, 2005) are applied to obtain estimates how the conditional probability $\mathbb{P}(y_n = 1|\tilde{\mathbf{x}}_n)$ depends on the explanatory variables $\tilde{\mathbf{x}}_n := (1, x_{n2}, x_{n3}, \ldots, x_{n7,1}, x_{nS,1}, \ldots, x_{nS,14})^\top \in \mathbb{R}^{k'}$, where $k' = 21$.

By means of the 1 as the first coordinate of $\tilde{\mathbf{x}}_n$, we include an intercept term. In addition, we abstract from feedback effects from $\tilde{\mathbf{x}}_n$ on y_n (in more technical terms we assume that the regressors $\tilde{\mathbf{x}}_n$ are exogenous; see, e.g., Davidson and MacKinnon, 1993, p. 624-627).

With probit and logit models $\mathbb{P}(y_n = 1|\tilde{\mathbf{x}}_n) = \mathbb{E}(y_n = 1|\tilde{\mathbf{x}}_n) = F(\boldsymbol{\beta}^\top \tilde{\mathbf{x}}_n)$, where $\boldsymbol{\beta} = (\beta_0, \beta_2, \ldots, \beta_{k'})^\top \in \mathbb{R}^{k'}$. The regression parameter β_i describes the impact of $\tilde{x}_{ni}$, i.e. the ith coordinate of $\tilde{\mathbf{x}}_n$, on the conditional probability $\mathbb{P}(y_n = 1|\tilde{\mathbf{x}}_n)$ (equal to the conditional expectation $\mathbb{E}(y_n = 1|\tilde{\mathbf{x}}_n)$), for $i = 0, 2, \ldots, k' = 21$, while $F(\cdot)$ is called link function. For the logit model the link function is provided by the logistic function, i.e.

$$\mathbb{P}(y_n = 1|\tilde{\mathbf{x}}_n) = \frac{e^{\boldsymbol{\beta}^\top \tilde{\mathbf{x}}_n}}{1 + e^{\boldsymbol{\beta}^\top \tilde{\mathbf{x}}_n}} \text{ , while for the probit model } \mathbb{P}(y_n = 1|\tilde{\mathbf{x}}_n) = \Phi\left(\boldsymbol{\beta}^\top \tilde{\mathbf{x}}_n\right) , \tag{1}$$

where $\Phi(\cdot)$ abbreviates the distribution function of the standard normal distribution. In this article parameter estimates, denoted by $\hat{\boldsymbol{\beta}}$, of the parameter vector $\boldsymbol{\beta}$ are obtained by means of maximum likelihood estimation (by using the `glm` function contained in the `R` package `AER`). To investigate the question how $\tilde{x}_{ni}$ affects $\mathbb{P}(y_n = 1|\tilde{\mathbf{x}}_n)$, the marginal effects

$$\frac{\partial}{\partial \tilde{x}_{ni}} \mathbb{E}(y_n = 1|\tilde{\mathbf{x}}_n) = F(\boldsymbol{\beta}^\top \tilde{\mathbf{x}}_n)\beta_i , \tag{2}$$

can be obtained, for $i = 0, 2, 3, \ldots, k' = 21$ (see, e.g., Greene, 1997; Cameron and Trivedi, 2005). In contrast to the linear regression model, the marginal effects described in (2) depend on the value of $\tilde{\mathbf{x}}_n$ where (2) is evaluated. In the following analysis, the term ME_i abbreviates the marginal effect $\frac{\partial}{\partial \tilde{x}_{ni}} \mathbb{E}(y_n = 1|\tilde{\mathbf{x}}_n)$ evaluated at the expected value $\mathbb{E}(\tilde{\mathbf{x}}_n)$. We obtain an estimate of the marginal effect, $\widehat{ME}_i$, by replacing $\boldsymbol{\beta}$ and $\mathbb{E}(\tilde{\mathbf{x}}_n)$ by their finite sample analogs $\hat{\boldsymbol{\beta}}$ and $\bar{\tilde{\mathbf{x}}}_n = \frac{1}{N}\sum_{n=1}^{N} \tilde{\mathbf{x}}_n$.

In contrast to the assumption of exogenous regressors, some users might have decided to 'install safety software', to 'read terms and conditions carefully at every registration', etc. *after* they had been offended and *before* they had been interviewed (in which case regressor endogeneity arises). If there are serious concerns that the persons interviewed behaved in this way, instrumental variable estimation should be performed, where we claim that finding good instruments for the given regression is a difficult problem. Although we can neither verify nor exclude that some interviewees acted in this way, we already observed in Table 2 inconsistent answers for a small group of interviewees (where we already argued that this can be due to mis-reporting or to a change in the behavior after an offense). To avoid possible problems arising from data points with inconsistent responses, we excluded those 48 observations $\mathbf{x}'_n$ where an interviewee n reported $y_n = 1$ and $x_{n2} = 0$ (regressions where these observations are still included are provided in Appendix E).

Tables 3 and 4 provide the regression results. By looking at the p-values, we observe that the regression intercept and the variable $Frequency$ are highly statistically significant for both models. The higher the variable $Frequency$ the larger the risk of an offense on the Internet. By means of the marginal effect we observe that a rise in the variable $Frequency$ by an infinitesimal unit, increases the probability to be offended by approximately 11% times this infinitesimal unit. When applying a significance level of 5% the variables *Employment* and $Human\ Capital$ are statistically insignificant, while at the 10% significance level the variables *Employment* and $Human\ Capital$ are (almost) significant. In more detail, for the variable *Employment* the p-values for the logit and the probit model are approximately 11% and 14%, while for the education variable $Human\ Capital$ the p-values are 10.4% and 9.6%, respectively. Since higher *Employment* reduces the risk to be offended on a significance close to 10%, the regressions provide weak support for the learning arguments provided Talib et al. (2010). Higher education, measured by the variable $Human\ Capital$, interestingly raises to probability to be offended at a significance level close to 10%. The impacts of the variables Age, $Gender$, and $Inhabitants$ are statistically insignificant (when applying significance levels $\leq 10\%$). Finally, we investigate the impacts arising from the various $Security$ variables $x_{nS,j}$. For both the logit and the probit model, the variable $x_{nS,6}$, 'only communicate with persons known in real life' is significant at a 5% significance level. The other $x_{nS,j}$ are not statistically significant at significance levels $\leq 10\%$.

4 Conclusions

To prevent and reduce the risk of individuals to be offended on the Internet, more detailed information on the socio-demographic as well as the risk-awareness characteristics of the users with respect to Internet security becomes necessary. This study uses questionnaire data from 3,000 Austrian individuals, recently collected by Kirchner et al. (2015), to provide information on these issues. The sample used in this article, contains information on employment, education, age and the frequency of Internet usage.

First, by means of a cluster analysis we investigate the question regarding the groups of persons being offended on the Internet. The cluster analysis suggests that offended users be partitioned into four groups, which are: A mainly female group, with group members living in small cities. In this cluster employment and the level of education are high. The second group is mainly male, living as well in smaller cities. For this group employment and education are also high, but slightly lower than in the female group. The third group of offended users lives mainly in large cities, with socio-demographic characteristics close to the values observed in the total sample of 3,000 individuals. However, this group exhibits the smallest awareness with respect to Internet security. The fourth group contains young users.

Second, after having identified these groups, we analyze the question whether the characteristics of the users such as age and gender as well as various protection methods applied by the users increase or decrease the risk to be offended on the Internet. By means of probit and logit regressions and applying a 5% significance level, we observe that the frequency of using the net raises the conditional probability to be offended, while to communicate only to people known in real life diminishes the conditional probability to be offended on the Internet. Variables like age and gender turned out to be statistically insignificant.

Table 2: Results obtained from the Cluster Analysis.

	Cluster												
Variable	1	2	3	4	5	6	7	8	9	10	11	12	mean/SD
Attacked	0.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	0.000	0.157
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.364
Frequency	0.358	1.709	1.057	1.527	0.621	1.474	1.138	0.290	1.381	1.556	0.000	1.786	0.957
	0.624	0.457	0.726	0.519	0.601	0.595	0.630	0.461	0.697	0.577	0.000	0.426	0.763
Gender	1.000	0.671	0.420	0.736	0.000	0.923	0.015	0.097	0.595	0.481	0.500	0.929	0.497
	0.000	0.473	0.495	0.441	0.000	0.268	0.121	0.301	0.497	0.509	0.577	0.267	0.500
Age	38.974	25.633	18.825	33.161	36.731	33.701	34.374	43.452	33.476	18.037	25.750	35.357	34.229
	8.033	6.764	5.082	8.859	9.128	9.282	9.212	5.847	8.454	5.185	5.560	9.982	10.048
Inhabitants	1.880	4.590	2.141	2.619	2.170	2.380	2.207	3.710	4.905	2.926	5.000	1.286	2.360
	1.480	0.986	1.569	1.785	1.638	1.747	1.634	1.792	0.370	1.817	0.000	0.469	1.720
Employment	1.459	1.250	1.180	1.309	1.051	1.297	0.989	0.000	1.094	1.500	1.000	0.000	1.197
	0.535	0.565	0.719	0.516	0.230	0.622	0.184	0.000	0.588	1.000	1.000	0.000	0.496
Human Capital	3.780	4.000	2.090	4.010	3.883	4.010	3.949	3.419	4.171	2.037	3.250	3.643	3.762
	0.937	0.816	0.581	0.839	0.944	0.864	0.902	1.259	0.803	0.192	0.957	1.447	1.020
$x_{nS,1}$	0.128	0.000	0.634	1.000	0.598	0.758	0.714	0.867	0.000	0.630	1.000	0.714	0.674
	0.335	0.000	0.483	0.000	0.491	0.430	0.453	0.352	0.000	0.492	0.000	0.469	0.469
Members	581.000	79.000	212.000	719.000	894.000	194.000	203.000	31.000	42.000	27.000	4.000	14.000	3000.000

[a] Results obtained from the cluster analysis. Data set $\mathbf{X}$, $N = 3,000$ observations, $k = 8$ variables, $I = 12$ clusters and l_1-distances. For each variable the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 12$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

A Cyber-Crime and Cyber-Crime Research

By considering the historical development, “cyber-crime emerged from hacking. Fraud schemes in relation with Social Engineering and other criminal activities were gradually added and connected to the technical and craft skills of the early hackers” (see Kochheim, 2016). While *information security research* is engaged in the development of software to increase IT security, *cyber-crime research* is connected to criminology and other social sciences with the goal to prevent cyber-crime (see, e.g., Hartel et al., 2011). Hartel et al. (2011)[Section 2] define *crime science* as applying scientific methods to prevent and to detect disorder, particularly crime. Then, referring to Newman (2009), the authors define *cyber-crime* as “behaviour in which computers or networks are a tool, a target, or a place of criminal activity.” For guidelines to perform information and communication technology research see, e.g. Bailey et al. (2012).

In addition, cyber-crime can be divided into “cyber-crime in a narrower sense”, where offenses are committed by using the technologies of the Internet (e.g., illegal access to a computer system), and “cyber-crime in a broader sense” (see, e.g., Bundeskriminalamt, 2015, p. 17), where the Internet is used as communication medium for criminal activity (e.g., fraud, child pornography and the initiation of sexual contacts with minors). In this article we refer to the broader definition of cyber-crime.

B Further Information about the Data

The study of Kirchner et al. (2015) is based on two surveys: The first sample comprises data from the Austrian population with an age between 14 and 49 years. The second sample considers parents (both or one parent) of children aged 10 to 13 years. In order to create the basis for the surveys and focus groups, 8 interviews with experts of the IT-division of the Austrian Ministry of the Interior (BM.I) as well as police-attorneys have been conducted. During the expert-interviews the problems of using the social media and future challenges were discussed. The results of the expert-interviews were used to design the questionnaires.

To obtain these data, Computer Assisted Telephone Interviews were performed. The data finally consists of 3,000 Austrians aged 14 to 49 years and 500 parents of children aged 10 to 13 years by using a standardized questionnaire. According to the requirements of the study, the characteristics of gender, age and place of residence (federal state) were considered as representative criteria. To obtain these

Table 3: Results obtained from the Logit Regression.

Variable	$\hat{\beta}_i$	SE	z-value	p-value	$\widehat{ME}_i$
Intercept	-2.8598	0.4772	-5.9920	0.0000	-0.4527
Frequency	0.7009	0.1147	6.1110	0.0000	0.1109
Gender	-0.1737	0.1348	-1.2890	0.1974	-0.0275
Age	0.0032	0.0077	0.4190	0.6749	0.0005
Inhabitants	0.0042	0.0379	0.1100	0.9121	0.0007
Employment	-0.2263	0.1409	-1.6060	0.1083	-0.0358
Human Capital	0.1157	0.0711	1.6270	0.1038	0.0183
Security $x_{n9,1}$	-0.1235	0.1658	-0.7450	0.4564	-0.0196
Security $x_{n9,2}$	0.0334	0.1374	0.2430	0.8077	0.0053
Security $x_{n9,3}$	0.0984	0.1585	0.6210	0.5349	0.0156
Security $x_{n9,4}$	0.2384	0.1725	1.3820	0.1670	0.0377
Security $x_{n9,5}$	0.0720	0.1447	0.4980	0.6187	0.0114
Security $x_{n9,6}$	-0.3298	0.1550	-2.1280	0.0334	-0.0522
Security $x_{n9,7}$	0.1061	0.1611	0.6580	0.5102	0.0168
Security $x_{n9,8}$	0.2231	0.1402	1.5920	0.1114	0.0353
Security $x_{n9,9}$	0.0941	0.1441	0.6530	0.5134	0.0149
Security $x_{n9,10}$	0.1036	0.1438	0.7200	0.4712	0.0164
Security $x_{n9,11}$	-0.1682	0.1639	-1.0260	0.3047	-0.0266
Incertitude $x_{n9,12}$	-0.0541	0.6816	-0.0790	0.9367	-0.0086
Security $x_{n9,13}$	-0.8227	0.7442	-1.1050	0.2689	-0.1302
Incertitude $x_{n9,14}$	-1.0265	0.7633	-1.3450	0.1787	-0.1625

[a] Results obtained from the logit regression. $\tilde{N} = N - 48 = 2,952$ observations, $1,708$ observations used by `R` due to missing values. y_n, i.e. 'personally offended', is the dependent variable, while $x_{n2}, \ldots, x_{nS,14}$ are the dependent variables. The second column provides the maximum likelihood estimates $\hat{\beta}_i$, $i = 0, 2, \ldots, k' = 21$, while the third, the forth and the fifth columns provide standard errors, z-values and p-values for the corresponding parameter estimates. A p-value of 0.000 denotes a p-value smaller than 0.0001. The last column shows estimates of the marginal effects ME_i.

Table 4: Results obtained from the Probit Regression.

Variable	$\hat{\beta}_i$	SE	z-value	p-value	$\widehat{ME}_i$
Intercept	-1.6943	0.2661	-6.3680	0.0000	-0.4757
Frequency	0.4102	0.0640	6.4080	0.0000	0.1152
Gender	-0.1045	0.0765	-1.3650	0.1721	-0.0293
Age	0.0016	0.0043	0.3720	0.7096	0.0005
Inhabitants	0.0018	0.0216	0.0840	0.9334	0.0005
Employment	-0.1170	0.0791	-1.4780	0.1394	-0.0328
Human Capital	0.0667	0.0401	1.6630	0.0963	0.0187
Security $x_{n9,1}$	-0.0684	0.0937	-0.7290	0.4658	-0.0192
Security $x_{n9,2}$	0.0118	0.0781	0.1510	0.8802	0.0033
Security $x_{n9,3}$	0.0633	0.0891	0.7100	0.4776	0.0178
Security $x_{n9,4}$	0.1342	0.0959	1.4000	0.1616	0.0377
Security $x_{n9,5}$	0.0417	0.0821	0.5080	0.6116	0.0117
Security $x_{n9,6}$	-0.1832	0.0892	-2.0530	0.0400	-0.0514
Security $x_{n9,7}$	0.0683	0.0914	0.7470	0.4549	0.0192
Security $x_{n9,8}$	0.1252	0.0800	1.5660	0.1173	0.0352
Security $x_{n9,9}$	0.0507	0.0817	0.6210	0.5349	0.0142
Security $x_{n9,10}$	0.0469	0.0812	0.5770	0.5638	0.0132
Security $x_{n9,11}$	-0.1000	0.0926	-1.0800	0.2801	-0.0281
Incertitude $x_{n9,12}$	0.0236	0.3903	0.0600	0.9519	0.0066
Security $x_{n9,13}$	-0.4284	0.3663	-1.1700	0.2422	-0.1203
Incertitude $x_{n9,14}$	-0.5229	0.3735	-1.4000	0.1615	-0.1468

[a] Results obtained from the probit regression. $\tilde{N} = N - 48 = 2,952$ observations, $1,708$ observations used by `R` due to missing values. y_n, i.e. 'personally offended', is the dependent variable, while $x_{n2}, \ldots, x_{nS,14}$ are the dependent variables. The second column provides the maximum likelihood estimates $\hat{\beta}_i$, $i = 0, 2, \ldots, k' = 21$, while the third, the forth and the fifth column provide standard errors, z-values and p-values for the corresponding parameter estimates. A *p*-value of 0.000 denotes a p-value smaller than 0.0001. The last column shows estimates of the marginal effects ME_i.

data, in total, about 50,000 people were contacted in order to achieve the desired 3,500 interviews. This corresponds to a response rate of around 7%. For about 37% of the calls, no one picked up; at about 18% the number from the phone book was invalid. Approximately 22% refused to participate in the survey and approximately 4% broke off the interview during the conversation.

The $N = 3,000$ survey was held in the period from July 9, 2014 to October 12, 2014. Some summary statistics are provided in Table 5. With the goal to obtain information on young users, in addition to the $N = 3,000$ sample used in this article, Kirchner et al. (2015) interviewed 500 parent(s) from December 11, 2014 until May 1, 2015. In those cases where the parents had more than one child in this age group, they were asked at the beginning of the interview how many children in this age group they have - and a random selection was set to which of their children they should refer.

For the sample of $N = 3,000$ interviews we observe the following: Let ζ stand for some attribute of the population measured in percentage terms. Then, given some point estimate $\hat{\zeta}$ based on the sample $\mathbf{X}$ of size $N = 3,000$, the 95% confidence interval (based on the normal approximation following from the asymptotic analysis) is $\left[\hat{\zeta} - 1.8\%, \hat{\zeta} + 1.8\%\right]$. In addition, by comparing the percentages observed for the population (third column in Table 5) to their sample analogs (fifth column in Table 5), we observe that all percentages observed for the population are contained in the interval "value observed in the sample $\pm$ standard error". By this we consider the survey samples as representative. That is, the distribution of the characteristics of gender, age and place of residence in the sample corresponds to that in the population.

Table 5: Sample vs. Population.

	Population		Sample	
	number	%	number	%
Total	4,043,432	100	3,000	100
Gender				
Male	2,035,814	50.35	1,510	50.33
Female	2,007,618	49.65	1,490	49.67
Age				
14-19 years	480,555	11.88	357	11.90
20-29 years	1,091,205	26.99	810	27.00
30-39 years	1,106,193	27.36	820	27.33
40-49 years	1,365,479	33.77	1,013	33.77
Provinces				
Burgenland	127,751	3.16	95	3.17
Carinthia	250,234	6.19	186	6.20
Lower Austria	746,531	18.46	554	18.47
Upper Austria	673,590	16.66	500	16.67
Salzburg	255,065	6.31	189	6.30
Styria	572,111	14.15	425	14.17
Tirol	353,340	8.74	262	8.73
Vorarlberg	181,905	4.50	135	4.50
Vienna	882,905	21.84	655	21.83

[a] The second columns presents the total number of individuals with an age between 14 and 49 years in Austria in the year 2014. The third column presents the percentages of the corresponding subgroups of the population. The forth column shows the number of individuals contained in the corresponding subgroup in the sample of $N = 3,000$ individuals. The last column presents the corresponding percentages.

Table 6: Descriptive Statistics - Security/Incertitude Variables $x_{nS,j}$.

Variable	$x_{nS,1}$	$x_{nS,2}$	$x_{nS,3}$	$x_{nS,4}$	$x_{nS,5}$	$x_{nS,6}$	$x_{nS,7}$	$x_{nS,8}$	$x_{nS,9}$	$x_{nS,10}$	$x_{nS,11}$	$x_{nS,12}$	$x_{nS,13}$	$x_{nS,14}$
Mean	0.674	0.420	0.627	0.731	0.507	0.726	0.716	0.333	0.552	0.583	0.221	0.006	0.016	0.020
SD	0.469	0.494	0.484	0.443	0.500	0.446	0.451	0.472	0.497	0.493	0.415	0.080	0.127	0.139
							Correlation							
$x_{nS,1}$	1.000	0.218	0.392	0.292	0.325	0.296	0.241	0.233	0.323	0.275	0.173	-0.001	-0.070	-0.204
$x_{nS,2}$	0.218	1.000	0.265	0.196	0.239	0.115	0.156	0.186	0.193	0.183	0.135	-0.003	-0.034	-0.121
$x_{nS,3}$	0.392	0.265	1.000	0.226	0.290	0.210	0.238	0.205	0.311	0.235	0.162	0.029	-0.041	-0.184
$x_{nS,4}$	0.292	0.196	0.226	1.000	0.274	0.191	0.240	0.183	0.236	0.283	0.086	-0.024	-0.083	-0.235
$x_{nS,5}$	0.325	0.239	0.290	0.274	1.000	0.245	0.293	0.240	0.253	0.248	0.174	-0.006	-0.036	-0.144
$x_{nS,6}$	0.296	0.115	0.210	0.191	0.245	1.000	0.304	0.127	0.226	0.208	0.103	-0.023	-0.065	-0.232
$x_{nS,7}$	0.241	0.156	0.238	0.240	0.293	0.304	1.000	0.151	0.240	0.191	0.102	-0.033	-0.054	-0.226
$x_{nS,8}$	0.233	0.186	0.205	0.183	0.240	0.127	0.151	1.000	0.181	0.195	0.120	0.011	-0.026	-0.101
$x_{nS,9}$	0.323	0.193	0.311	0.236	0.253	0.226	0.240	0.181	1.000	0.240	0.154	-0.003	-0.048	-0.158
$x_{nS,10}$	0.275	0.183	0.235	0.283	0.248	0.208	0.191	0.195	0.240	1.000	0.138	0.025	-0.070	-0.168
$x_{nS,11}$	0.173	0.135	0.162	0.086	0.174	0.103	0.102	0.120	0.154	0.138	1.000	-0.017	-0.020	-0.076
$x_{nS,12}$	-0.001	-0.003	0.029	-0.024	-0.006	-0.023	-0.033	0.011	-0.003	0.025	-0.017	1.000	-0.010	-0.011
$x_{nS,13}$	-0.070	-0.034	-0.041	-0.083	-0.036	-0.065	-0.054	-0.026	-0.048	-0.070	-0.020	-0.010	1.000	-0.018
$x_{nS,14}$	-0.204	-0.121	-0.184	-0.235	-0.144	-0.232	-0.226	-0.101	-0.158	-0.168	-0.076	-0.011	-0.018	1.000

[a] Descriptive Statistics Variable $x_{nS,j}$, $N = 3,000 - 681 = 2319$ observations. *Mean* abbreviates the sample mean, *SD* the sample standard deviation and *Correlation* for the Pearson correlation.

C Agglomerate Hierarchical Clustering

By means of a cluster analysis we try to find groups within a data set (the following section is mainly based on Kaufman and Rousseeuw, 1990, Chapters 2, 3 and 5). The data consists of N observations $\mathbf{x}_n \in \mathbb{R}^k$, $n = 1, \ldots, N$, where k is the dimension of column vector $\mathbf{x}_n$. $\mathbb{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ stands for the data set, which can also be written in terms of the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times k}$. In particular, our data set consists of $N = 3,000$ individuals who filled in the questionnaire, while $k = 8$ is the number of attributes taken from the questionnaire. x_{n1} and x_{nS} in this section corresponds to y_n and $x_{nS,1}$ in the main text.

Since the data are measured on different scales, the standardized observations $z_{ni} := \frac{x_{ni} - \hat{\mu}_i}{\widehat{sd}_i}$ are often used, when a cluster analysis is performed. $\hat{\mu}_i := \frac{1}{N} \sum_{n=1}^{N} x_{ni}$ stands for the sample mean and $\widehat{sd}_i := \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (x_{ni} - \hat{\mu}_i)^2}$ stands for the sample standard deviation of attribute i, where $i = 1, \ldots, k$. We also follow this approach and standardize the observations x_{ni}, resulting in $z_{ni} \in \mathbb{R}$ and $\mathbf{z}_n := (z_{n1}, \ldots, z_{nk})^\top \in \mathbb{R}^k$.

To measure the degree of dissimilarity between the observations $\mathbf{z}_n$ and $\mathbf{z}_m$ a distance function $d(\cdot, \cdot)$ has to be chosen (if the data are not standardized, replace $\mathbf{z}_n$ and $\mathbf{z}_m$ by $\mathbf{x}_n$ and $\mathbf{x}_m$). In the following we work with l_1-distances (= Manhattan distances in `R`)

$$\mathtt{d}_1(\mathbf{z}_n, \mathbf{z}_m) := \sum_{i=1}^{k} |z_{ni} - z_{mi}| \ , \tag{3}$$

as well as with Euclidean distances

$$\mathtt{d}_2(\mathbf{z}_n, \mathbf{z}_m) := \sqrt{\sum_{i=1}^{k} (z_{ni} - z_{mi})^2} \ . \tag{4}$$

After having defined distances between observations $\mathbf{z}_n$ and $\mathbf{z}_m$, we want to obtain distances between some clusters $\mathscr{C}_i$ and $\mathscr{C}_j$. A cluster $\mathscr{C}_i$ is a subset of $\mathbb{X}$, where $\mathscr{C}_1, \ldots, \mathscr{C}_I$ partition the set $\mathbb{X}$. That is $\mathscr{C}_i \neq \emptyset$, $\mathscr{C}_i \cap \mathscr{C}_j = \emptyset$ for all $i, j = 1, \ldots, I$, where $i \neq j$, and $\bigcup_{i=1}^{I} \mathscr{C}_i = \mathbb{X}$. I stands for the number of clusters considered. Equipped with the definition of $\mathscr{C}_i$, we define the distance between $\mathscr{C}_i$ and $\mathscr{C}_j$ as follows:

$$d_v\left(\mathscr{C}_i, \mathscr{C}_j\right) := \frac{1}{|\mathscr{C}_i||\mathscr{C}_j|} \sum_{n\in\mathscr{C}_i m\in\mathscr{C}_j} \mathtt{d}_v(\mathbf{z}_n, \mathbf{z}_m) \ , \tag{5}$$

where $v \in \{1, 2\}$. $|\mathscr{C}_i|$ and $|\mathscr{C}_j|$ stand for the number of elements of the sets $\mathscr{C}_i$ and $\mathscr{C}_j$. Literature calls the distance defined in (5) "unweighted pair-group average method".

As already stated in the main text, the *Agglomerate hierarchical clustering* technique `agnes` described in Kaufman and Rousseeuw (1990)[Chapter 5] is applied in our study. Agglomerate hierarchical clustering techniques start with N clusters, that is $\mathscr{C}_n = \{\mathbf{x}_n\}$ for $n = 1, \ldots, N$, and then merge the groups according to the value of the distance function. This procedure is continued until we end up with $\mathscr{C}_1 = \mathbb{X}$ and $|\mathscr{C}_1| = N$. Differences in various agglomerate hierarchical clustering methods are mainly due to differences in the distance measures.

In more detail, in this study we proceed as follows: Let I_ℓ stand for the number of clusters in step ℓ, $\mathscr{C}_{i,\ell}$, where $i = 1, \ldots, I_\ell$, for the clusters obtained in step ℓ, $d_v\left(\mathscr{C}_{i,\ell}, \mathscr{C}_{j,\ell}\right)$ for the corresponding distances between $\mathscr{C}_{i,\ell}$ and $\mathscr{C}_{j,\ell}$ and $d_{v,[\ell,1]}(\mathscr{C}_{q,\ell}, \mathscr{C}_{w,\ell})$ for the smallest distance between $\mathscr{C}_{i,\ell}$ and $\mathscr{C}_{j,\ell}$, where $j, i = 1, \ldots, I_\ell$ and $i \neq j$, in step ℓ. Let the pair with the smallest distance have the indexes q and w, where $q, w \in \{1, \ldots, I_\ell\}$. Table 7 demonstrates how an agglomerate hierarchical clustering algorithm starts with N clusters, where $\mathscr{C}_{n,\ell=0} = \{\mathbf{x}_n\}$, and ends up with one cluster $\mathscr{C}_{1,\ell=N-1} = \mathbb{X}$ in the final step. To obtain the distance between the clusters, the data are standardized. Then Euclidean and l_1-distances are applied in (5) to obtain the distances between the clusters (see equation (5)). As described in Kaufman and Rousseeuw (1990)[page 205] the "dissimilarity between merging clusters" is monotone. In more formal terms, $d_{v,[\ell,1]}(\mathscr{C}_{q,\ell}, \mathscr{C}_{w,\ell}) \geq d_{v,[\ell-1,1]}(\mathscr{C}_{q,\ell-1}, \mathscr{C}_{w,\ell-1})$ for $\ell = 0, \ldots, N-2$. By collecting these dissimilarities we obtain the monotone increasing sequence of "levels" $l_0 := d_{v,[\ell=0,1]}(\mathscr{C}_{q,0}, \mathscr{C}_{w,0}) \leq l_1 := d_{v,[\ell=1,1]}(\mathscr{C}_{q,1}, \mathscr{C}_{w,1}) \leq \cdots \leq l_{N-2} := d_{v,[\ell=N-1,1]}(\mathscr{C}_{q,N-1}, \mathscr{C}_{w,N-2})$.

By considering the step $\ell = \mathtt{h}$, where observation $\mathbf{x}_n$ is merged the first time with some $\mathscr{C}_w$, we observe the dissimilarity $d_{v,[\mathtt{h}-1,1]}(\mathbf{x}_n, \mathscr{C}_{w,\mathtt{h}-1}) = g_n$ at this merger. By calculating g_n/l_{N-2} we obtain a number in the interval $[0, 1]$. g_n/l_{N-2} is often called "width of the banner n", since the factions g_n/l_{N-2} can be presented in terms of a banner plot. Kaufman and Rousseeuw (1990)[page 211] interpret g_n/l_{N-2} as "... it gives an idea of the amount of structure that has been found by the algorithm. Indeed, when the data

Table 7: Agglomerate hierarchical clustering algorithm.

`Step 0:`	$\mathscr{C}_{n,0} = \{\mathbf{x}_n\}$, for $n = 1, \ldots, N = I_0$.
`Step 1:`	Take the distances $d_v\left(\mathscr{C}_{i,0}, \mathscr{C}_{j,0}\right)$, where $i, j \in \{1, \ldots, I_0\}$,
	obtain $d_{v,[0,1]}(\mathscr{C}_{q,0}, \mathscr{C}_{w,0})$,
	merge $\mathscr{C}_{q,0}$ and $\mathscr{C}_{w,0}$, i.e. $\mathscr{C}_{q,1} = \mathscr{C}_{q,0} \cup \mathscr{C}_{w,0}$.
$\vdots$	$\vdots$
`Step` ℓ:	Take the distances $d_v\left(\mathscr{C}_{i,\ell-1}, \mathscr{C}_{j,\ell-1}\right)$, where $i, j \in \{1, \ldots, I_{\ell-1}\}$,
	obtain $d_{v,[\ell-1,1]}(\mathscr{C}_{q,\ell-1}, \mathscr{C}_{w,\ell-1})$,
	merge $\mathscr{C}_{q,\ell-1}$ and $\mathscr{C}_{w,\ell-1}$, i.e. $\mathscr{C}_{q',\ell} = \mathscr{C}_{q,\ell-1} \cup \mathscr{C}_{w,\ell-1}$.
$\vdots$	$\vdots$
`Step N-1:`	Take the distances $d_v\left(\mathscr{C}_{i,N-2}, \mathscr{C}_{j,N-2}\right)$, where $i, j \in \{1, \ldots, I_{N-2}\}$,
	obtain $d_{v,[N-2,1]}(\mathscr{C}_{q,N-2}, \mathscr{C}_{w,N-2})$,
	merge $\mathscr{C}_{1,N-2}$ and $\mathscr{C}_{2,N-2}$, i.e. $\mathscr{C}_{1,N-1} = \mathscr{C}_{1,N-2} \cup \mathscr{C}_{2,N-2} = \mathbb{X}$.

[a] The distances between clusters $d_v(\mathscr{C}_{q,\ell}, \mathscr{C}_{w,\ell})$ follow from (5), the distances between the standardized observations from (3) and (4

possess a clear cluster structure, the between-cluster dissimilarities and hence the highest level (l_{N-2} in our notation) will become much larger than the within-cluster dissimilarities, and as a consequence the black lines become longer ($1 - g_n/l_{N-2}$ becomes larger in our notation).” The mean of these fractions

$$AC := \frac{1}{N} \sum_{n=1}^{N} \frac{g_n}{l_{N-2}} \tag{6}$$

is called *agglomerative coefficient*. The higher AC the better the explanatory power of the cluster analysis.

The *dendrogram* (clustering tree) is a graphical representation of the results obtained by a hierarchical clustering technique. On the vertical axis the observations indices $n = 1, \ldots, N$ are arranged, such that the branches of the tree do not intersect. On the vertical axis we observe the levels l_n. The corresponding branches describe the leaves of the tree to be merged. The “height” of a branch represents the difference – in terms of levels – between the corresponding groups $\mathscr{C}_q$ and $\mathscr{C}_w$ to be merged. In formal terms, the heights are obtained by means of $\frac{1}{|\mathscr{C}_q|!} \sum_{n,m \in \mathscr{C}_q} d_v(\mathbf{x}_n, \mathbf{x}_m) - \frac{1}{|\mathscr{C}_w|!} \sum_{n,m \in \mathscr{C}_w} d_v(\mathbf{x}_n, \mathbf{x}_m)$. The dendrograms

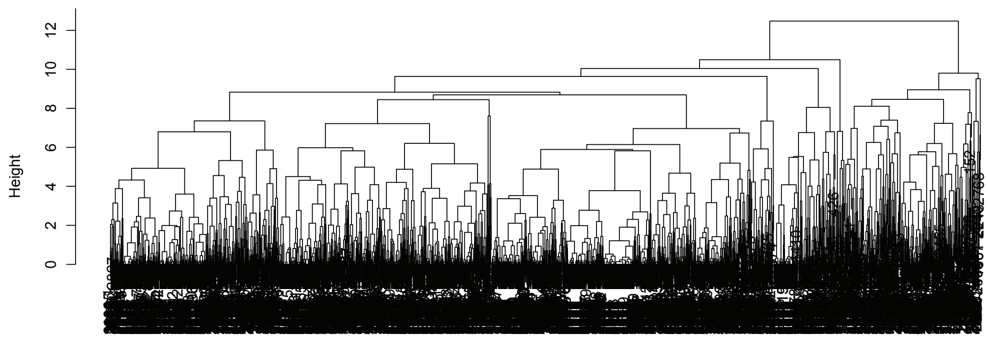

Figure 1: Dendrogram for Internet Security data This figure plots the dendrogram for the Internet security data $\mathbf{X}$. $N = 3,000$ and $k = 8$. l_1 distances are applied here. Heights on the vertical axis, individuals arranged according to the tree structure on the horizontal axis. Agglomerative Coefficient $AC = 0.98$.

for the data set $\mathbf{X}$ are provided in the Figures 1 and 2, for l_1 and Euclidean distances, respectively. E.g., in Figure 1 we observe the final transition from two groups to one group at a height of 12. At a height close to 8 we observe already twelve groups, etc.

The agglomerate hierarchical clustering algorithm `agnes` was implemented in the `R`-package in Maechler et al. (2015). In particular, our estimates are obtained by means of the R-commands: `agnes`($\mathbf{X}$,`diss` = `FALSE`, `metric` = `manhattan`, `stand` = `TRUE`, `method` = `average`) for l_1-distances. For Euclidean distances set `metric` = `euclidian`. `stand` = `TRUE` means that the data are standardized, `method` = `average` implies that unweighted pair-group averages are used. For more details see Maechler et al. (2015) and the literature cited in this manual.

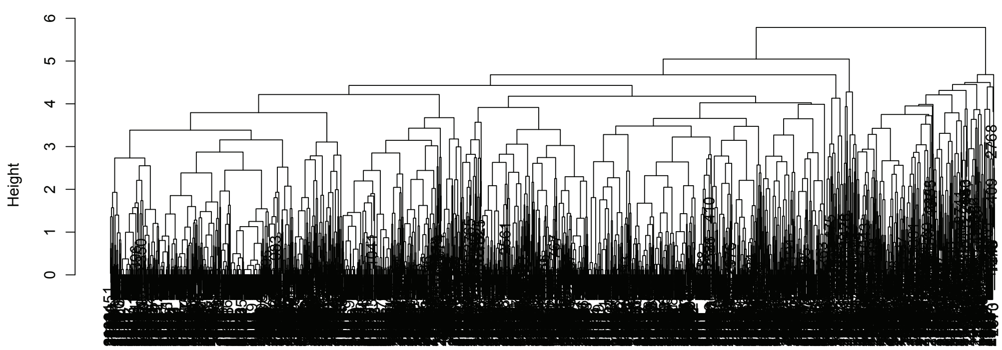

Figure 2: Dendrogram for Internet Security data This figure plots the dendrogram for the Internet security data $\mathbf{X}$. $N = 3,000$ and $k = 8$. Euclidean distances are applied here. Heights on the vertical axis, individuals arranged according to the tree structure on the horizontal axis. Agglomerative Coefficient $AC = 0.94$.

D Further Clustering Results

This section provides further clustering results, with different numbers of clusters I as well as clustering results with Euclidean distances (4). NA denotes a static which is not available. In the following tables this takes place for the sample standard deviation when the number of group members is one.

With $I = 4$ classes all $\mathcal{O} = 470$ persons subject to an attack are contained in the class $\mathscr{C}_3$ with l_1-distances (3), while with Euclidean distances the classes $\mathscr{C}_2$ and $\mathscr{C}_4$ contain offended users (see Tables 8 and 9).

With $I = 8$ clusters we observe that the class $\mathscr{C}_3$ in Table 8 splits up into the classes $\mathscr{C}_4$, $\mathscr{C}_6$, $\mathscr{C}_7$ and $\mathscr{C}_8$ in Table 10. For Euclidean distances the class $\mathscr{C}_2$ in Table 9 splits up into the classes $\mathscr{C}_3$, $\mathscr{C}_5$ and $\mathscr{C}_7$ in Table 11, while the class $\mathscr{C}_4$ with $I = 4$ is now labeled $\mathscr{C}_8$.

With $I = 12$ classes, the group $\mathscr{C}_4$ splits up into $\mathscr{C}_6$ and $\mathscr{C}_7$ in Table 10, while the former classes $\mathscr{C}_6$, $\mathscr{C}_7$ and $\mathscr{C}_8$ are labeled $\mathscr{C}_9$, $\mathscr{C}_{10}$ and $\mathscr{C}_{11}$ in Table 2. For Euclidean distances we observe that the classes $\mathscr{C}_5$ and $\mathscr{C}_8$ remain the same, the new labels with with $I = 12$ are $\mathscr{C}_7$ and $\mathscr{C}_{11}$. The class $\mathscr{C}_3$ splits up into the classes $\mathscr{C}_4$ and $\mathscr{C}_{10}$, while $\mathscr{C}_7$ splits up into the classes $\mathscr{C}_9$ and $\mathscr{C}_{12}$ in Table 12.

With $I = 16$ groups, the classes $\mathscr{C}_7$, $\mathscr{C}_{10}$ and $\mathscr{C}_{11}$ remain the same, with the new labels $\mathscr{C}_7$, $\mathscr{C}_{13}$ and $\mathscr{C}_{14}$. The former class $\mathscr{C}_6$ splits up into $\mathscr{C}_6$ and $\mathscr{C}_8$, while $\mathscr{C}_9$ splits up into $\mathscr{C}_{11}$ and $\mathscr{C}_{12}$ in Table 13. For Euclidean distances only the labeling is changed for the classes $\mathscr{C}_7$, $\mathscr{C}_9$, $\mathscr{C}_{10}$, $\mathscr{C}_{11}$ and $\mathscr{C}_{12}$, i.e. these classes are $\mathscr{C}_8$, $\mathscr{C}_{13}$, $\mathscr{C}_{14}$, $\mathscr{C}_{15}$ and $\mathscr{C}_{16}$ in Table 14. Finally, class $\mathscr{C}_4$ in Table 12 splits up into the classes $\mathscr{C}_5$ and $\mathscr{C}_{10}$ in Table 12.

Table 8: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, l_1-distances and $I = 4$ clusters.

	Cluster				
Variable	1	2	3	4	mean/SD
Attacked	0.000	0.000	1.000	0.000	0.157
	0.000	0.000	0.000	0.000	0.364
Frequency	0.884	1.057	1.313	0.290	0.957
	0.767	0.726	0.648	0.461	0.763
Gender	0.514	0.420	0.472	0.097	0.497
	0.500	0.495	0.500	0.301	0.500
Age	35.785	18.825	33.004	43.452	34.229
	9.185	5.082	9.712	5.847	10.048
Inhabitants	2.315	2.141	2.585	3.710	2.360
	1.706	1.569	1.794	1.792	1.720
Employment	1.229	1.180	1.129	0.000	1.197
	0.468	0.719	0.489	0.000	0.496
Human Capital	3.899	2.090	3.877	3.419	3.762
	0.914	0.581	0.969	1.259	1.020
$x_{nS,1}$	0.679	0.634	0.666	0.867	0.674
	0.467	0.483	0.472	0.352	0.469
Members	2287.000	212.000	470.000	31.000	3000.000

[a] For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 4$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

Table 9: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, Euclidean distances and $I = 4$ clusters.

	Cluster				
Variable	1	2	3	4	mean/SD
Attacked	0.000	1.000	0.000	1.000	0.157
	0.000	0.000	0.000	0.000	0.364
Frequency	0.898	1.306	0.489	1.833	0.957
	0.765	0.648	0.626	0.408	0.763
Gender	0.505	0.472	0.311	0.500	0.497
	0.500	0.500	0.468	0.548	0.500
Age	34.320	32.858	42.022	44.333	34.229
	10.091	9.676	7.031	4.676	10.048
Inhabitants	2.293	2.595	3.750	1.833	2.360
	1.691	1.796	1.780	1.602	1.720
Employment	1.235	1.148	0.000	0.000	1.197
	0.469	0.470	0.000	0.000	0.496
Human Capital	3.748	3.891	3.378	2.833	3.762
	1.022	0.967	1.230	0.408	1.020
$x_{nS,1}$	0.674	0.662	0.778	1.000	0.674
	0.469	0.474	0.424	0.000	0.469
Members	2485.000	464.000	45.000	6.000	3000.000

[a] Euclidean distances. For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 4$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

Table 10: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, l_1-distances and $I = 8$ clusters.

	Cluster								
Variable	1	2	3	4	5	6	7	8	mean/SD
Attacked	0.000	0.000	0.000	1.000	0.000	1.000	1.000	1.000	0.157
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.364
Frequency	0.854	1.709	1.057	1.302	0.290	1.381	1.556	0.000	0.957
	0.759	0.457	0.726	0.635	0.461	0.697	0.577	0.000	0.763
Gender	0.509	0.671	0.420	0.458	0.097	0.595	0.481	0.500	0.497
	0.500	0.473	0.495	0.499	0.301	0.497	0.509	0.577	0.500
Age	36.149	25.633	18.825	34.045	43.452	33.476	18.037	25.750	34.229
	9.053	6.764	5.082	9.240	5.847	8.454	5.185	5.560	10.048
Inhabitants	2.234	4.590	2.141	2.291	3.710	4.905	2.926	5.000	2.360
	1.670	0.986	1.569	1.690	1.792	0.370	1.817	0.000	1.720
Employment	1.228	1.250	1.180	1.129	0.000	1.094	1.500	1.000	1.197
	0.465	0.565	0.719	0.466	0.000	0.588	1.000	1.000	0.496
Human Capital	3.896	4.000	2.090	3.980	3.419	4.171	2.037	3.250	3.762
	0.917	0.816	0.581	0.883	1.259	0.803	0.192	0.957	1.020
$x_{nS,1}$	0.713	0.000	0.634	0.736	0.867	0.000	0.630	1.000	0.674
	0.453	0.000	0.483	0.442	0.352	0.000	0.492	0.000	0.469
Members	2208.000	79.000	212.000	397.000	31.000	42.000	27.000	4.000	3000.000

[a] For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 8$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

Table 11: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, Euclidean distances and $I = 8$ clusters.

	Cluster								
Variable	1	2	3	4	5	6	7	8	mean/SD
Attacked	0.000	0.000	1.000	0.000	1.000	0.000	1.000	1.000	0.157
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.364
Frequency	0.379	1.419	1.263	0.489	1.455	0.532	1.563	1.833	0.957
	0.587	0.539	0.638	0.626	0.689	0.620	0.619	0.408	0.763
Gender	0.490	0.524	0.446	0.311	0.545	0.387	0.656	0.500	0.497
	0.500	0.500	0.498	0.468	0.503	0.491	0.483	0.548	0.500
Age	39.783	29.726	34.528	42.022	29.491	20.661	18.969	44.333	34.229
	7.296	9.677	8.972	7.031	9.426	4.428	3.729	4.676	10.048
Inhabitants	1.913	2.582	2.205	3.750	4.891	3.917	3.219	1.833	2.360
	1.486	1.782	1.652	1.780	0.369	1.488	1.791	1.602	1.720
Employment	1.335	1.111	1.174	0.000	0.971	1.214	0.875	0.000	1.197
	0.478	0.415	0.442	0.000	0.568	0.738	0.835	0.000	0.496
Human Capital	3.882	3.667	4.016	3.378	3.852	2.790	2.500	2.833	3.762
	0.970	1.039	0.876	1.230	1.035	0.994	0.762	0.408	1.020
$x_{nS,1}$	0.513	0.770	0.735	0.778	0.000	0.067	0.938	1.000	0.674
	0.500	0.421	0.442	0.424	0.000	0.252	0.246	0.000	0.469
Members	1191.000	1232.000	377.000	45.000	55.000	62.000	32.000	6.000	3000.000

[a] For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 8$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

Table 12: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, Euclidean distances and $I = 12$ clusters.

	Cluster												
Variable	1	2	3	4	5	6	7	8	9	10	11	12	mean/SD
Attacked	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	1.000	1.000	1.000	1.000	0.157
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	NA	0.364
Frequency	0.304	1.419	0.523	1.271	0.625	0.414	1.455	0.532	1.613	1.105	1.833	0.000	0.957
	0.535	0.539	0.654	0.641	0.806	0.501	0.689	0.620	0.558	0.567	0.408	NA	0.763
Gender	0.285	0.524	0.889	0.422	0.250	0.345	0.545	0.387	0.677	0.895	0.500	0.000	0.497
	0.452	0.500	0.315	0.495	0.447	0.484	0.503	0.491	0.475	0.315	0.548	NA	0.500
Age	40.159	29.726	39.054	34.148	46.313	39.655	29.491	20.661	18.677	41.684	44.333	28.000	34.229
	7.513	9.677	6.805	8.969	3.361	7.437	9.426	4.428	3.400	5.386	4.676	NA	10.048
Inhabitants	1.990	2.582	1.764	2.174	1.333	5.000	4.891	3.917	3.161	2.833	1.833	5.000	2.360
	1.538	1.782	1.370	1.640	0.488	0.000	0.369	1.488	1.791	1.823	1.602	NA	1.720
Employment	0.997	1.111	2.000	1.127	0.000	0.000	0.971	1.214	1.000	2.000	0.000	0.000	1.197
	0.075	0.415	0.000	0.406	0.000	0.000	0.568	0.738	0.816	0.000	0.000	NA	0.496
Human Capital	3.918	3.667	3.811	4.080	2.313	3.966	3.852	2.790	2.516	2.842	2.833	2.000	3.762
	0.992	1.039	0.922	0.844	0.793	1.017	1.035	0.994	0.769	0.602	0.408	NA	1.020
$x_{nS,1}$	0.401	0.770	0.702	0.732	0.556	0.889	0.000	0.067	0.935	0.789	1.000	1.000	0.674
	0.491	0.421	0.459	0.444	0.527	0.323	0.000	0.252	0.250	0.419	0.000	NA	0.469
Members	786.000	1232.000	405.000	358.000	16.000	29.000	55.000	62.000	31.000	19.000	6.000	1.000	3000.000

[a] For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 12$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

Table 13: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, l_1-distances and $I = 16$ clusters.

	Cluster																
Var.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	m./SD
Atta.	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	0.16
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	0.36
Freq.	0.36	1.71	1.38	1.53	0.62	1.28	1.14	1.54	0.29	0.00	1.37	2.00	1.56	0.00	1.57	2.00	0.96
	0.62	0.46	0.49	0.52	0.60	0.65	0.63	0.56	0.46	0.00	0.70	NA	0.58	0.00	0.53	0.00	0.76
Gender	1.00	0.67	0.45	0.74	0.00	0.96	0.01	0.91	0.10	0.32	0.61	0.00	0.48	0.50	1.00	0.86	0.50
	0.00	0.47	0.50	0.44	0.00	0.20	0.12	0.28	0.30	0.47	0.49	NA	0.51	0.58	0.00	0.38	0.50
Age	38.97	25.63	18.22	33.16	36.73	32.79	34.37	33.99	43.45	20.78	33.49	33.00	18.04	25.75	26.71	44.00	34.23
	8.03	6.76	4.68	8.86	9.13	10.35	9.21	8.93	5.85	5.85	8.56	NA	5.18	5.56	1.38	6.30	10.05
Inha.	1.88	4.59	2.06	2.62	2.17	1.13	2.21	2.79	3.71	2.42	4.95	3.00	2.93	5.00	1.57	1.00	2.36
	1.48	0.99	1.55	1.78	1.64	0.34	1.63	1.83	1.79	1.61	0.22	NA	1.82	0.00	0.53	0.00	1.72
Empl.	1.46	1.25	1.06	1.31	1.05	1.61	0.99	1.21	0.00	1.36	1.13	0.00	1.50	1.00	0.00	0.00	1.20
	0.53	0.56	0.75	0.52	0.23	0.50	0.18	0.63	0.00	0.64	0.56	NA	1.00	1.00	0.00	0.00	0.50
Hum.C.	3.78	4.00	1.95	4.01	3.88	3.98	3.95	4.02	3.42	2.54	4.15	5.00	2.04	3.25	5.00	2.29	3.76
	0.94	0.82	0.35	0.84	0.94	0.71	0.90	0.91	1.26	0.89	0.80	NA	0.19	0.96	0.00	0.49	1.02
$x_{nS,1}$	0.13	0.00	0.70	1.00	0.60	0.09	0.71	0.97	0.87	0.28	0.00	0.00	0.63	1.00	0.71	0.71	0.67
	0.34	0.00	0.46	0.00	0.49	0.28	0.45	0.16	0.35	0.46	0.00	NA	0.49	0.00	0.49	0.49	0.47
Memb.	581.000	79.000	162.000	719.000	894.000	47.000	203.000	147.000	31.000	50.000	41.000	1.000	27.000	4.000	7.000	7.000	3000.000

[a] For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 16$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

Table 14: Results obtained from clustering. Data $\mathbf{X}$, $N = 3,000$, $k = 8$, Euclidean distances and $I = 16$ clusters.

	Cluster																
Var.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	m./SD
Atta.	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.16
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	0.36
Freq.	0.30	1.44	1.42	0.52	1.33	0.63	0.41	1.45	1.36	1.11	0.47	0.63	1.61	1.11	1.83	0.00	0.96
	0.53	0.55	0.53	0.65	0.61	0.81	0.50	0.69	0.54	0.69	0.56	0.71	0.56	0.57	0.41	NA	0.76
Gend.	0.28	0.50	0.54	0.89	0.41	0.25	0.34	0.55	0.53	0.45	0.00	1.00	0.68	0.89	0.50	0.00	0.50
	0.45	0.50	0.50	0.31	0.49	0.45	0.48	0.50	0.50	0.50	0.00	0.00	0.48	0.32	0.55	NA	0.50
Age	40.16	31.22	29.40	39.05	33.60	46.31	39.66	29.49	27.62	35.65	21.89	18.71	18.68	41.68	44.33	28.00	34.23
	7.51	8.71	10.27	6.81	8.46	3.36	7.44	9.43	9.39	10.12	4.51	3.58	3.40	5.39	4.68	NA	10.05
Inha.	1.99	4.74	1.50	1.76	2.53	1.33	5.00	4.89	1.23	1.20	4.71	2.55	3.16	2.83	1.83	5.00	2.36
	1.54	0.54	1.06	1.37	1.75	0.49	0.00	0.37	0.47	0.59	0.57	1.60	1.79	1.82	1.60	NA	1.72
Empl.	1.00	1.25	0.98	2.00	1.09	0.00	0.00	0.97	1.15	1.23	1.17	1.30	1.00	2.00	0.00	0.00	1.20
	0.07	0.45	0.35	0.00	0.39	0.00	0.00	0.57	0.40	0.43	0.79	0.67	0.82	0.00	0.00	NA	0.50
Hum.C.	3.92	4.18	3.36	3.81	4.16	2.31	3.97	3.85	3.47	3.86	3.34	1.92	2.52	2.84	2.83	2.00	3.76
	0.99	0.85	1.01	0.92	0.82	0.79	1.02	1.04	1.08	0.88	0.85	0.41	0.77	0.60	0.41	NA	1.02
$x_{nS,1}$	0.40	0.84	1.00	0.70	1.00	0.56	0.89	0.00	0.00	0.00	0.11	0.00	0.94	0.79	1.00	1.00	0.67
	0.49	0.37	0.00	0.46	0.00	0.53	0.32	0.00	0.00	0.00	0.31	0.00	0.25	0.42	0.00	NA	0.47
Memb.	786.000	430.000	589.000	405.000	262.000	16.000	29.000	55.000	213.000	96.000	38.000	24.000	31.000	19.000	6.000	1.000	3000.000

[a] For each variable presented in the first column, the first row presents group-specific sample means in the corresponding cluster $\mathscr{C}_i$, $i = 1, \ldots, 16$, while the second row presents the group-specific sample standard deviations. The last column presents the mean values and the sample standard deviations for the corresponding variables, obtained from all observations $n = 1, \ldots, N$. The last row presents the number of individuals assigned to cluster $\mathscr{C}_i$.

E Further Regression Results

Tables 15 and 16 provide the regression results when all $N = 3,000$ observations are used and the 48 interviewees with inconsistent replies are not excluded. By looking at the p-values, we observe that the regression intercept is highly statistically significant for both models. In addition, the variable $Frequency$ becomes significant in both models if a significance level of approximately 6.5% is applied. The higher the variable $Frequency$ the larger the risk of an offense on the Internet. As already observed in the main text, the variables $Employment$ and $Human\ Capital$ are (almost) significant on a 10% significance level. The impacts of the variables Age, $Gender$, and $Inhabitants$ are statistically insignificant (when applying a significance level $\leq 10\%$).

Next, we investigate the impacts arising for the various $Security$ variables $x_{nS,j}$. For both the logit and the probit model, $x_{nS,4}$, 'install safety software', $x_{nS,8}$, 'read terms and conditions carefully at every registration', and $x_{nS,13}$, 'do not use social networks' are statistically significant at a 10% significance level, while the p-value for the variable $x_{nS,1}$, 'adapt protection settings at the first registration' is slightly above 10%. The variable $x_{nS,6}$, 'only communicate with persons known in real life' is significant at a 5% significance level. The other $x_{nS,j}$ are not statistically significant. When considering the signs of the (almost) significant variables $x_{nS,j}$, we observe that the signs of the parameter estimates, and thereby the marginal effects, are – as expected – negative for the variables $x_{nS,1}$, $x_{nS,6}$ and $x_{nS,13}$. However, for the variables $x_{nS,4}$ and $x_{nS,8}$ (both significant at the 10% level), we observe the counterintuitive result that these possible measures to increase protection raises the probability to be offended on the Internet.

Finally, we would like to remark on this counterintuitive result that the variables $x_{nS,4}$ and $x_{nS,8}$ have a positive and significant impact on the probability to be offended (at the 10% level). Positive parameter estimates $\hat{\beta}_4$ and $\hat{\beta}_8$ can be due to sampling effects, interviewees mis-understanding the questions, etc. In addition, as already stated in Section 3, in contrast to the assumption of exogenous regressors, the users might have decided to 'install safety software' and to 'read terms and conditions carefully at every registration' *after* they had been offended and *before* they had been interviewed. At least for the 48 interviewees where inconsistencies in x_{n2} and y_n are observed, we have concerns that that some users in this subgroup behaved in this way. Hence, the results with $\tilde{N} = N - 48$ interviewees are presented in the main text. In contrast to Tables 15 and 16, when excluding these observations, the impacts of the variables $x_{nS,4}$ and $x_{nS,8}$ are statistically insignificant (see, e.g., p-values in Tables 3 and 4).

Table 15: Results obtained from the Logit Regression.

Variable	$\hat{\beta}_i$	SE	z-value	p-value	$\widehat{ME}_i$
Intercept	-2.0506	0.4377	-4.6850	0.0000	-0.3246
Frequency	0.1843	0.0997	1.8490	0.0645	0.0292
Gender	-0.1557	0.1275	-1.2210	0.2221	-0.0246
Age	0.0021	0.0072	0.2930	0.7697	0.0003
Inhabitants	0.0262	0.0356	0.7350	0.4622	0.0041
Employment	-0.2132	0.1332	-1.6000	0.1096	-0.0337
Human Capital	0.1100	0.0667	1.6480	0.0993	0.0174
Security $x_{nS,1}$	-0.2505	0.1534	-1.6330	0.1024	-0.0435
Security $x_{nS,2}$	-0.0240	0.1306	-0.1840	0.8540	-0.0038
Security $x_{nS,3}$	0.1797	0.1489	1.2070	0.2276	0.0284
Security $x_{nS,4}$	0.3160	0.1635	1.9330	0.0533	0.0500
Security $x_{nS,5}$	0.0526	0.1372	0.3840	0.7013	0.0083
Security $x_{nS,6}$	-0.3437	0.1476	-2.3280	0.0199	-0.0544
Security $x_{nS,7}$	0.1083	0.1538	0.7040	0.4812	0.0171
Security $x_{nS,8}$	0.2604	0.1326	1.9640	0.0495	0.0412
Security $x_{nS,9}$	0.0563	0.1363	0.4130	0.6798	0.0089
Security $x_{nS,10}$	0.1344	0.1366	0.9840	0.3253	0.0213
Security $x_{nS,11}$	-0.1556	0.1556	-1.0000	0.3171	-0.0246
Incertitude $x_{nS,12}$	0.0129	0.6742	0.0190	0.9847	0.0020
Security $x_{nS,13}$	-1.2771	0.7385	-1.7290	0.0838	-0.2022
Incertitude $x_{nS,14}$	-0.1976	0.5203	-0.3800	0.7041	-0.0313

[a] Results obtained from the logit regression. $N = 3,000$ observations, $1,749$ observations used by `R` due to missing values. y_n, i.e. 'personally offended', is the dependent variable, while $x_{n2}, \ldots, x_{nS,14}$ are the dependent variables. The second column provides the maximum likelihood estimates $\hat{\beta}_i$, $i = 0, 2, \ldots, k' = 21$, while the third, the forth and the fifth columns provide standard errors, z-values and p-values for the corresponding parameter estimates. A p-value of 0.000 denotes a p-value smaller than 0.0001. The last column shows estimates of the marginal effects ME_i.

Table 16: Results obtained from the Probit Regression.

Variable	$\hat{\beta}_i$	SE	z-value	p-value	$\widehat{ME}_i$
$Intercept$	-1.2337	0.2483	-4.9690	0.0000	-0.3464
$Frequency$	0.1074	0.0569	1.8880	0.0590	0.0302
$Gender$	-0.0933	0.0732	-1.2750	0.2024	-0.0262
Age	0.0013	0.0041	0.3160	0.7518	0.0004
$Inhabitants$	0.0149	0.0205	0.7270	0.4669	0.0042
$Employment$	-0.1169	0.0757	-1.5440	0.1225	-0.0328
$Human\ Capital$	0.0642	0.0381	1.6860	0.0917	0.0180
$Security\ x_{nS,1}$	-0.1440	0.0882	-1.6320	0.1027	-0.0404
$Security\ x_{nS,2}$	-0.0187	0.0751	-0.2490	0.8037	-0.0052
$Security\ x_{nS,3}$	0.1075	0.0848	1.2670	0.2050	0.0302
$Security\ x_{nS,4}$	0.1809	0.0918	1.9700	0.0488	0.0508
$Security\ x_{nS,5}$	0.0331	0.0788	0.4200	0.6745	0.0093
$Security\ x_{nS,6}$	-0.1952	0.0859	-2.2720	0.0231	-0.0548
$Security\ x_{nS,7}$	0.0663	0.0881	0.7530	0.4517	0.0186
$Security\ x_{nS,8}$	0.1494	0.0766	1.9510	0.0511	0.0419
$Security\ x_{nS,9}$	0.0284	0.0783	0.3630	0.7167	0.0080
$Security\ x_{nS,10}$	0.0714	0.0780	0.9160	0.3599	0.0200
$Security\ x_{nS,11}$	-0.0889	0.0888	-1.0010	0.3170	-0.0250
$Incertitude\ x_{nS,12}$	0.0352	0.3894	0.0900	0.9280	0.0099
$Security\ x_{nS,13}$	-0.6657	0.3532	-1.8850	0.0595	-0.1869
$Incertitude\ x_{nS,14}$	-0.1041	0.2836	-0.3670	0.7136	-0.0292

[a] Results obtained from the probit regression. $N = 3,000$ observations, $1,749$ observations used by `R` due to missing values. y_n, i.e. 'personally offended', is the dependent variable, while $x_{n2}, \ldots, x_{nS,14}$ are the dependent variables. The second column provides the maximum likelihood estimates $\hat{\beta}_i$, $i = 0, 2, \ldots, k' = 21$, while the third, the forth and the fifth column provide standard errors, z-values and p-values for the corresponding parameter estimates. A p-value of 0.000 denotes a p-value smaller than 0.0001. The last column shows estimates of the marginal effects ME_i.

References

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., Moore, T., and Savage, S. (2013). Measuring the cost of cybercrime. In *The economics of information security and privacy*, pages 265–300. Springer Berlin Heidelberg.

Bailey, M., Dittrich, D., Kenneally, E., and Maughan, D. (2012). The menlo report. *IEEE Security Privacy*, 10(2):71–75.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76:1–54.

BM.I (2015). Bundesministerium für Inneres (BM.I; Federal Ministry of the Interior, Austria); Sicherheitsbericht 2014. `http://www.bmi.gv.at/cms/BK/presse/files/KrimBericht_2014_web.pdf`. Accessed: 2015-10-20.

Bullée, J.-W. H., Montoya, L., Pieters, W., Junger, M., and Hartel, P. H. (2015). The persuasion and security awareness experiment: reducing the success of social engineering attacks. *Journal of Experimental Criminology*, 11(1):97–115.

Bundeskriminalamt (2015). Sicherheit 2014 (.BK; Federal Criminal Police Office, Austria). `http://www.bmi.gv.at/cms/BK/presse/files/KrimBericht_2014_web.pdf`. Accessed: 2015-10-20.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Cook, P. J., Machin, S., Marie, O., and Mastrobuoni, G. (2014). Lessons from the economics of crime. CentrePiece - The Magazine for Economic Performance 410, Centre for Economic Performance, LSE.

CPSR (2015). Computer Professionals for Social Responsibility: The Ten Commandments of Computer Ethics. `http://cpsr.org/issues/ethics/cei/`. Accessed: 2015-11-24.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Dimkov, T. (2012). *Alignment of organizational security policies : theory and practice*. PhD thesis, Enschede, the Netherlands. IPA Dissertation Series no. 2012-04.

E.C. (2016). European Commission: Digital Agenda for Europe. `http://ec.europa.eu/digital-agenda/`. Accessed: 2016-01-04.

Freeman, R. (1999). *The Economics of Crime*, page Chapter 52. North Holland Publishers, Amsterdam, Netherlands.

Greene, W. H. (1997). *Econometric Analysis*. Perentice Hall, New Jersey, 3rd edition.

Halaweh, M. and Fidler, C. (2008). Security perception in e-commerce: Conflict between customer and organizational perspectives. In *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, pages 443–449.

Hartel, P., Junger, M., and Wieringa, R. (2011). Cyber-crime science = crime science + information security. Technical Report TR-CTI, University of Enschede, Enschede.

Hinduja, S. and Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, 29(2):129–156.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley.

Kirchner, S., Angleitner, B., and Gstrein, M. (2015). Cyber Crime - die Social Media-Nutzer in Österreich und ihre Erfahrungen mit kriminalpolizeilich relevanten Aktivitäten. Technical report, Institute for Advanced Studies, Vienna and MAKAM Research GmbH; this project was financed by the Security Research Program KIRAS of the Austrian Ministry for Transport, Innovation and Technology.

Kochheim, D. (2016). Modulares Cybercrime . `http://www.cyberfahnder.de/doc/Kochheim,%20Modulares%20Cybercrime.pdf`. Accessed: 2016-09-02.

Kshetri, N. (2010). *The Global Cybercrime Industry: Economic, Institutional and Strategic Perspectives*. Springer Berlin Heidelberg.

Lee, M., Crofts, T., Salter, M., Milivojevic, S., and McGovern, A. (2013). 'let's get sexting': Risk, power, sex and criminalisation in the moral domain. *International Journal for Crime, Justice and Social Democracy*, 2(1):35–49.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3.

Newman, G. R. (2009). *Cybercrime*, pages 551–584. Springer New York, New York, NY.

OeIAT (2016). Österreichisches Institut für angewandte Telekommunikation (ÖIAT). `https://www.saferinternet.at`. Accessed: 2016-01-04.

Schneider, C., Katzer, K., and Leest, U. (2013). Cyberlife – Spannungsfeld zwischen Faszination und Gefahr: Cybermobbing bei Schülerinnen und Schülern - Eine empirische Bestandsaufnahme bei Eltern, Lehrkräften und Schülern/innen in Deutschland. Technical report, Bündnis gegen Cybermobbing e.V., Leopoldstr. 1, 76133 Karlsruhe.

Statistik Austria (2015). IKT-Einsatz in Haushalten 2015. `http://www.statistik.at/`. Accessed: 2016-07-25.

Talib, S., Clarke, N. L., and Furnell, S. M. (2010). An analysis of information security awareness within home and work environments. In *Availability, Reliability, and Security, 2010. ARES '10 International Conference on*, pages 196–203.

Tsohou, A., Kokolakis, S., Karyda, M., and Kiountouzis, E. (2008). Investigating information security awareness: Research and practice gaps. *Information Security Journal: A Global Perspective*, 17(5-6):207–227.