Private and Public Information Disclosure to Improve Cybersecurity

A field experiment to incentivise compliance with anti-spoofing best-practices

Master Thesis Report

by

Luigi Tuttobene

TU Delft

Private and Public Information Disclosure to Improve Cybersecurity

A field experiment to incentivise compliance with anti-spoofing best-practices

Master thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE
in **Management of Technology**
Faculty of Technology, Policy and Management

by
Luigi Tuttobene
Student number: 4522729

To be defended in public on May 24th, 2018

Graduation committee

Chairperson: Prof. M.J.G. van Eeten, Section Organisation & Governance
First Supervisor: Dr. C. H. Gañán, Section Organisation & Governance
Second Supervisor: Dr. G.A. de Reuver, Section Information and Communication Technology
Advisor: O. Çetin, Section Organisation & Governance

Executive summary

This thesis deals with the notification of cybersecurity vulnerabilities. Given the critical importance of the Internet in our society, engaging individual and organisation in better security behaviours has become a major challenge for researchers and policy makers fighting cybercrime. In particular, finding effective ways to report security vulnerability and to instigate remediation represents a frontline in this fight. In fact, detecting possible points of failure of the Internet is not enough, if the security issue is not adequately addressed. Unfortunately, there are many obstacles in the process of reporting security vulnerabilities to the affected party. From identifying the responsible party to retrieving the right contact information, from the content of the message to the selection of a means of communication, many things can, and do go wrong.

On top of these operational aspects of vulnerability notification, there are additional motivational issues that contribute to complicate this process. In fact, as remediation costs time, effort and money, actors may lack sufficient *incentives* to act on abuse notifications, especially when they are not directly affected by the lack of security. Arguably, misaligned incentives between who is responsible for security and who enjoys its benefits (or who suffers its lack) are a prominent cause of security failures: when the costs of in-security do not encumber on the vulnerable party, but are instead internalised by other actors or society at large, that party has no incentives to remediate.

Thus, cybersecurity researchers have turn their attention to design and test policy interventions to cope with these incentives problems. One debated approach refers to the use of public disclosure policies to stimulate additional incentives out of reputation concern and peer pressure. In fact, it has been argued that publicly revealing information about people or organisations' wrongdoing represents a viable way to prompt compliance with norms and regulations. The same logic might apply to cybersecurity: publicly notifying vulnerable parties by openly revealing their lack of security may be functional to induce public disapproval, which in turn can instigate compliance with security practices. However, public notifications are not only about *naming and shaming*: disclosing relevant information also carries a message about appropriate and desirable behaviours, contributing to raise awareness about the issues that requires regulatory attention.

In this thesis, we focus on a particular security problem, IP address spoofing. Despite being a well-known issue for more than 30 years, IP spoofing remains a popular attack method, due to a misalignment in the incentives of the actors involved in the remediation of the problem. IP spoofing is the illicit practice of creating Internet packets with a fake source address, to hide the real sender identity or to impersonate another computer network. Network operators must comply with anti-spoofing best practices (i.e. deploying *source address validation* filters) to prevent attackers from launching massive distributed deny of service attacks. Note, however, that deploying anti-spoofing filters does not directly increase the security of the network deploying it: operators can follow all best practices, still receive anonymous, malicious traffic from other operators who do not have proper filters. Moreover, additional difficulties in measuring compliance with anti-spoofing and the lack of enforceable regulations further reduce operators' incentives to deploy.

Over the years, researchers have collected measurements about which networks are compliant and which lack proper anti-spoofing measures. In our research, we seek to aggregate these results and use them as targeted feedback for operators, in form of private and public notification. In fact, we hypothesise that disclosing information about network lacking anti-spoofing might incentivise operators to remediate. Therefore, we formulate the following research question: *to what extent do notifications incentivise compliance with anti-spoofing best practices?*

To answer this question, we conduct a field quasi experiment aimed at testing the effect of privately and publicly disclose information about which network are compliant and which lack anti-spoofing. In particular, operators found without anti-spoofing are assigned to three experimental groups: a control group (which receive no treatment), a private disclosure group (for which information is shared only to the operators involved), and a public disclosure group (for which this information is also shared with a selection of third parties, including CERTs, Network Operator Groups and security bloggers). To disclose compliance information, we design Infospoofing.com, a website on which we regularly release measurements of compliance. On the website we list operators lacking anti-spoofing as well as general statistics on its adoption. To better manage the disclosure, our intervention is at a country level: "spoofable" operators are grouped by country, and countries are assigned to the experimental conditions. Despite true experimental designs require random assignment, in our case complete randomisation might be problematic, as we want to compare the impact of our intervention on groups composed out of similar countries. Thus, groups of similar countries are formed by means of a cluster analysis on the basis of GDP per capita, ICT Development Index and Global Cybersecurity Index. In this way, we select three triplets of countries (one for treatment), for a total of 99 autonomous systems.

In this setting, we notified via an email the operators of 67 networks (30 in the private group and 37 in the public). The message included the IP addresses that showed evidences of spoofing, the link to our website, and country level statistics on the deployment of anti-spoofing. The difference between private and public group is that in the public group we also share our website with third parties, so that differences in remediation are due to the "*publicity effect*".

Of the 67 operators notified, 27 opened the link to our website (40.3%). This is already a positive result: previous studies that included a demonstrative website got very little engagement. We attribute this fact to the nudging tone of our notification: since the number of networks found without anti-spoofing is a little part of the total number of tested networks, the tone of the notification was crafted accordingly, to highlight such disproportion. Moreover, we received 12 automatic acknowledgements and 7 manual replies. We took the 7 manual replies as a chance to further investigate the factors that prevent operators from deploying. It has emerged that a couple reported technical limitations of anti-spoofing filters (increase fragility of the network or is not compatible with current infrastructure). Interestingly, most of these operators were already deploying anti-spoofing on other network segments, and were not aware about the existence of the problem.

As for the promotion of the website, we successfully engaged the operators' attention via the Network Operator Groups. Some operators reached out to us, interested in our project. It looks like the public

disclosure of vulnerability information might bring positive side effects on the community, increasing awareness and establishing a culture of best practices.

Next, we analysed the remediation rate. We observed remediation in 10 cases on 67 (14.9%). In particular, 4 ASes on 30 (13.3%) remediated in the private group, and 6 on 37 (16.2%) in the public group. No evidence of remediation has been observed in the control group.
To investigate differences the remediation rates of each group, survival analysis is performed. Our results show that both private and public notifications have a positive impact on remediation, as the probability of remediation in the private and public groups are significantly different from the control group. However, our analysis failed to identify significant differences between private and public notifications in terms of remediation rate. These results suggest that is the notification itself, rather that the type of notification, that has an impact.
Finally, we use logistic regression analysis to test whether the occurrence of remediation can be modelled as a function of operators' organisational factors (i.e. the type of network notified, the size of the network), socio-technical characteristics of the country (i.e. GDP per capita, ICT Development Index and Global Cybersecurity Index) and on the visits to our website. The results show that only the visit to our website is a significant predictor of remediation, meaning that visiting our website increases the likelihood of deploying filters.

All in all, our conclusion is that notifying operators has a moderate effect, still positive, on operators' incentives to deploy anti-spoofing filters.

The private notification is affected by problems in reaching the affected party. The contact information we used were retrieved by WHOIS look up, a standard query-response protocol that provide information about the owner or the responsible of Internet resources like IP addresses. Unfortunately, in many case the information provided is obsolete or incorrect, because operators do not regularly update it. The flip side is that, even if these operators received the notification, they ignored it, showing additional lack of care. For this reason, we encourage future research to engage in dialogue with these operators, and to better understand their incentives.

The public disclosure of information keeps appearing a viable solution, in light of the good engagement reached via NOG. In this regard, it is notably that our treatment spilled over from the public group to private: at some point, our website has been posted on the NOG of a country in the private group. Though, looking at the results, it does not seem that this had interfered on the success of the experiment, it has highlighted that the community of operators is active and sympathetic to the problem. For this reason, we suggest using NOGs to keep posting monthly report of recent measurement of anti-spoofing compliance, trying to observe whether, in the long run, this might produce additional benefits.

Table of contents

Executive summary i
Table of contents iv
List of Figures vii
List of Tables vii
Chapter 1: Research proposal 1
1.1 Introduction 1
1.2 Research objectives 2
1.2.1 Knowledge gap 3
1.2.2 Research objective 3
1.2.3 Research questions 4
1.3 Research approach 5
1.4 Contributions 6
1.4.1 Scientific relevance 6
1.4.2 Deliverables 7
Chapter 2: An economical insight on IP address spoofing 9
2.1 The Economics of Cybersecurity 9
2.1.1 Externalities 10
2.1.2 Asymmetric information 12
2.1.3 Misaligned incentives 12
2.1.4 Proposed policy solutions 13
2.2 Economics of IP spoofing 16
2.2.1 IP source address spoofing 17
2.2.2 Reflection and amplification DDoS attack 17
2.2.3 Ingress filtering and source address validation 20
2.2.4 Measuring deployment of source address validation 21
2.2.5 Incentives analysis 23
2.3 Conclusions 31
Chapter 3: A framework for public notifications 32
3.1 Sharing security information 32
3.2 Vulnerability notifications 33
3.3 Social information and public disclosure 36
3.3.1 Social comparison and pro-social behaviour 36
3.3.2 Public disclosure as a mean of compliance 37

3.3.3 Final thoughts 39
3.4 Conceptual framework 40
3.4.1 Formulating the research question 40
3.4.2 Conceptual model 42
3.5 Conclusions 45
Chapter 4: Research methodology 46
4.1 Overview on the experimental design 46
4.2 Data gathering and aggregation 47
4.2.1 List of spoofable ASes 47
4.2.2 Infospoofing.com 48
4.3 Pre-test and post-test measurements 51
4.4 Country selection and assignment 52
4.4.1 Country selection 52
4.4.2 Cluster analysis 54
4.4.3 Country assignment 54
4.5 Pre-test crowdsourcing measurements 56
4.5.1 Pre-test results 56
4.6 Treatments 57
4.6.1 Experimental groups 57
4.6.2 Notifications to non-compliant operators 58
4.6.3 Notifications to third parties 58
4.7 Measuring remediation 59
4.7.1 Intention to remediate 59
4.7.2 Deployment of filters 60
4.7.3 Defining remediation 61
4.8 Putting all together 62
4.9 Conclusions 64
Chapter 5: Results and data analysis 65
5.1 Notification analysis 65
5.1.1 Notifications to operators 65
5.1.2 Analysis of operators' reaction 66
5.1.3 Third party engagement 67
5.1.4 Increase in spoofable networks observed 69
5.2 Post-test measurements and remediation 70

5.3 Data analysis 71
5.3.1 Survival analysis 71
5.3.2 Regression analysis 76
5.4 Discussion 79
5.4.1 Hypothesis validation 79
5.4.2 Discussion of the results 80
5.5 Conclusions 82
Chapter 6: Discussion and conclusions 83
6.1 Answering the research questions 83
6.1.1 Reviewing the research questions 83
6.1.2 Private notification 83
6.1.3 Public notification 84
6.1.4 Role of operators' characteristics 85
6.1.5 Recommendations 85
6.1.6 Private and public disclosure to improve cybersecurity? 88
6.2 Limitation 88
6.3 Future research 89
References 90
Appendix 1: Interpreting results of the Spoofer test 99
Appendix 2: Preliminary interviews 105
1. Aim of the interviews 105
2. Methodology 105
2.1 Approach 105
2.2 Interview guide 107
2.3 Sample 108
3. Analysis 109
3.1 Interpretation 109
3.2 Results 109
4. Limitations 114
Appendix 3: Cluster analysis 115
Appendix 4: Crowdsourcing measurements 117
1. Measurement infrastructure 117
2. Results 119
Appendix 5: Notification to non-compliant operators 121
Appendix 6: Questionnaire to non-compliant operators 122

Reporting IP spoofing 122
Appendix 7: Notification for NOGs 124
Appendix 8: Notification for national CERTs 125
Appendix 9: Notification to security blogs 126
Appendix 10: Logistic analysis on the visits to the website 127
Appendix 11: R code for logistic analysis 129

List of Figures

Figure 1. Conceptual model. 6
Figure 2. Research structure. 8
Figure 3. Schematization of a reflection and amplification DDoS attack 18
Figure 4. Schematisation of source address validation 20
Figure 5. Details on the Spoofer application (adapted from Beverly & Bauer, 2007) 23
Figure 6. Main factors contributing to operators' incentives (adapted from van Eeten et al., 2010). . 25
Figure 7. Conceptual model. 44
Figure 8. Homepage of our website 50
Figure 9. Website page showing compliance information in Germany 51
Figure 10. Dendrogram (left) and elbow method (right). 55
Figure 11. Visualization of the countries selected for the experiment grouped by cluster 55
Figure 12. Experimental groups 57
Figure 13. Review of the experimental design. 63
Figure 14. Survival probability for all ASes (no cluster distinction). 73
Figure 15. Survival probability (first cluster). 74
Figure 16. Survival probability (second cluster). 74
Figure 17. Survival probability (third cluster). 74
Figure 18. Survival probability for all ASes grouped by cluster.. 74
Figure 19. Survival probability for all ASes who visited the website. 74
Figure 20. Model diagnosis with ROC curve. 78
Figure 21. Tests collected from the prefix 213.152.165.0/24 100
Figure 22. Details of test results. 101
Figure 23. Test collected from the prefix 195.8.192.0/24. 102
Figure 24. Test collected from AS20860. 102
Figure 25. Evidence of remediation from the prefix 180.214.68.0/24. 104
Figure 26. Scatter plot, correlation matrix and histogram. 115
Figure 27. Structure of the crowdsourcing measurements. 117
Figure 28. Model diagnosis with ROC curve. 128

List of Tables

Table 1. Network operators' incentives to deploy filters. 29

Table 2. List of countries most covered by Prolific. 53
Table 3. Experimental groups 56
Table 4. Final number of ASes included in the experiment. 57
Table 5. Notification results. 66
Table 6. NOGs visit to our website grouped by page visited. 68
Table 7. Visits to our website grouped by page visited. 68
Table 8. Number of unique IP addresses that visited our website grouped by IP country of origin 69
Table 9. Increase in spoofable ASes after our experiment. 69
Table 10. Remediation. 71
Table 11. Log rank test results (no cluster distinction) 73
Table 12. Log rank test results for the first cluster. 75
Table 13. Log rank test results for the second cluster. 75
Table 14. Log rank test results for the third cluster. 75
Table 15. Log rank test results (grouped by cluster). 75
Table 16. Results of logistic regression analysis. 78
Table 17. Goodness-of-fit 78
Table 18. Crowdsourcing pre-test results. 119
Table 19. Results of logistic model for visit to the website 128
Table 20. Goodness-of-fit of the model 128

Chapter 1:
Research proposal

1.1 Introduction

The Internet has not been designed keeping security in mind. Many security mechanism, like encryption and authentication, were not part of its original protocol suite. Rather, they have been introduced and incorporated in the Internet as an afterthought (Anderson, 2010). As a result, the implementation of important defensive mechanisms depends on the willingness of individual actors to invest in security. Security measures that do not yield a substantial return on investments are likely to face a hard time before being widely adopted, though their rapid diffusion is beneficial for the whole Internet ecosystem (Anderson & Moore, 2006). Thus, in spite of an initial technological facade, most of the challenges in cybersecurity can be reframed in light of motivational issues: technological solutions exist, but actors lack sufficient incentives to adopt them (Anderson, 2001). Consequently, researchers have highlighted the need of designing strategies to tackle these *incentive problems*, and to instigate security behaviours that are desirable from the point of view of the entire Internet ecosystem (Anderson & Moore, 2007; van Eeten & Bauer, 2008; Anderson, Böhme, Clayton & Moore, 2009; Moore, 2010).

Research aimed at instigating similar *pro-social behaviours* has recently focused on conditional cooperation: individual contributions to a common good are higher when information is provided that many others are contributing (Frey & Meier, 2004). Empirical findings have shown the positive effects of disclosing such information in the case of charitable donations (Frey & Meier, 2004; Shang & Croson, 2009), political participation (Margetts, Escher & Reissfelder, 2011), household electricity and water savings (Grønhøj & Thøgersen, 2011; Ferraro & Price, 2013), contribution to online communities (Butler, 2001; Ludford, Cosley, Frankowski & Terveen, 2004) and many other fields.
In addition, policies aimed at disclosing relevant social information have been a valuable tool for policy makers to prompt organisations to comply with norms and regulations, albeit their success is conditional to many variables. When effective, disclosure policies are: *"A magic cocktail of instrumental utility and social meaning"* (Kahan, 2006). Nevertheless, the strength of that cocktail varies according to the nature of the information disclosed, stakeholders' reactions and the authority and legitimacy of the disclosing party (Pawson, 2002; Kahan, 2006; Hutter & Jones, 2007; Lee, 2010; van Erp, 2011).

Whether disclosure policies might fit cybersecurity problems is the object of investigation of this research. In particular, our focus is on the use of private and public information disclosure to increase the compliance with anti-spoofing best practices. IP address spoofing, a fundamental problem in Internet architecture and root cause of massive DDoS attacks, is the practice of forging part of the header of Internet packets, alternating the source IP address to mask sender's identity or to impersonate another computer system (Internet Society, 2015). It exploits a design problem in Internet's architecture: traffic is forwarded taking only care of the destination address, the source address is not

validated. So far, measures to implement *source address validation* are formulated in terms of best current practices: network operators should filter their traffic to prevent Internet packets with a non-verified address to leave their network (Ferguson & Senie, 2000; Baker & Savola, 2004). In practice, the compliance with SAV is *incentive misaligned* (as operators can follow all best practices and still receive anonymous, malicious traffic from third-parties who do not properly filter), and *incentive incompatible* (as not compliant operators have no intention to reveal their lack of compliance) (Beverly, Berger & Hyun, 2009).

In an attempt to infer the extent of anti-spoofing filters on the Internet, researchers of the Center for Applied Internet Data Analysis (CAIDA) launched the Spoofer Project, providing volunteers with a software to test the presence of anti-spoofing filters of their network (Beverly & Bauer, 2005). Over the years, their initiative collected measurements of SAV compliance from more than 3,000 networks (Beverly, Koga & Claffy, 2013), and the results are publicly available on the project website (CAIDA, 2018). According to these measurements, around 30% of the networks tested appears, to some extent, *spoofable*[1] (i.e. the test revealed the lack of anti-spoofing filter at least on an IP prefix of the network). When the Spoofer test reveals the lack of SAV on a network, researchers of CAIDA report to the operator of that network. So far, these notifications induced an encouraging remediation rate swinging between 15% and 20% (Claffy, 2017).
However, given the increasing rate and volume of attacks based on IP spoofing, an important question to address is: *how can we further incentivise compliance with anti-spoofing best practices?*

This research investigates the use of information disclosure to notify non-compliant network operators, and to incentivise them to deploy anti-spoofing filters out of reputation concern and peer pressure. To be more precise, we seek to understand which factors determine operator's incentives to deploy filters, and, in particular, which factors prevent them from doing so. Next, we shall test the effectiveness of notifying non-compliant operators by privately disclosing information about which networks are already SAV compliant and which are instead lacking anti-spoofing filters. Moreover, we are interested in understanding whether the public disclosure of this information may engage the network operator community and other relevant third parties, and thus generating additional pressure on non-compliant operators.

The rest of this proposal is organised as follows: Section 1.2 formulates research objectives and research questions, Section 1.3 describes the research approach, and Section 1.4 outlines the main contribution of this research.

1.2 Research objectives

This section presents the knowledge gap, the research objectives, and the questions this research seeks to address.

[1] Note that throughout this research we will refer to a network found without anti-spoofing as *spoofable*, as also done by researchers of CAIDA. To refer to the operators of the spoofable networks we will use the term *non-compliant operators* (referring to the fact that they do not comply with anti-spoofing best practices).

1.2.1 Knowledge gap

Given the critical importance of the Internet in our society, engaging individual and organisation in better security behaviours has become a major challenge for researchers and policy makers fighting cybercrime. In particular, finding effective ways to report security vulnerability and to instigate remediation represents a frontline in this fight (Jhaveri, Cetin, Gañán, Moore, & van Eeten, 2017). In fact, detecting possible points of failure of the Internet is not enough, if the security issue is not adequately addressed. However, as remediation costs, actors may lack sufficient incentives to act on vulnerability notifications.
In this research, we seek to understand whether disclosure policies can solicit additional incentives out of reputation concern and peer pressure. Such type of *public notification* has been around for a while (Arora, Telang & Xu, 2004; Moore & Clayton 2011, He, Lee, Han & Whinston, 2016), and contributed to create an interesting debate: should vulnerability information be publicly disclosed? Supporters of public disclosure argue that it further encourages remediation, whereas opponents of vulnerability disclosure argue that it provides attackers with information they may not obtain otherwise.

The hypothesis in this research is that public notification may be an effective way to incentivise network operators to comply with anti-spoofing best practices. Despite being a well-known security problem for more than 30 years (Morris, 1985), IP address spoofing remains a popular attack vector, as evidenced in March 2018 during a 1.7 Tbps attack confirmed by Arbor Networks (Morales, 2018). Previous attempts to shed light on the problem focus mainly on the technological component, in order to design defences to prevent attacks using IP spoofing (Ferguson & Senie, 2000; Baker & Savola, 2004), or to detect networks not deploying such defences (Beverly & Bauer, 2005, 2007; Lone, Luckie, Korczyński & van Eeten, 2017). Nevertheless, this technical branch of the literature has acknowledged that the fundamental obstacle to the widespread adoption of anti-spoofing measures is a lack of incentives (Beverly & Bauer, 2007; Internet Society, 2015). Authors pinpoint a set of factors that might cause this problem, including the law individual benefits, lack of business care, costs of deployment and technical limitations. However, it remains unclear how operators perceive these factors, and, in general, how compelling they value the mitigation of IP spoofing.

Thus, the knowledge gap addressed in this thesis refers to the incentives of network operators to deploy anti-spoofing filters, and, in particular, on the role of reputation and peer pressure on such incentives.

1.2.2 Research objective

In light of the knowledge gap just described, the objective of this research is:

to study the effect of private and public notifications on the compliance with anti-spoofing best practices.

In order to achieve this objective, the following complementary goals are formulated:

- To investigate the problem of IP spoofing, and to identify the factors that determine operators' incentives to deploy anti-spoofing measures;
- To design a field experiment to test the effect of private and public notifications on the compliance with anti-spoofing best practices;

- To identify practical recommendations for policy makers and cybersecurity researchers.

1.2.3 Research questions

Having formulated the research objectives, in this section we present the research question we seek to answer (further discussed in Section 3.3.1). The principal question in this research is:

- *RQ: To what extent do notifications incentivise compliance with anti-spoofing best practices?*

This question represents the main focus of the research, and it is aimed at investigating the effectiveness of private and public notifications at getting operators deploy anti-spoofing filters. In order to answer this main research question, additional research sub-questions are formulated:

1. *SQ1: What is the effect of privately notifying non-compliant operators?*

First, we seek to understand the effect of private notifications on the deployment of anti-spoofing. We shall notify non-compliant operators, bringing to their attention that their network has been found without anti-spoofing filters and demanding remediation. To make the message more compelling, the notification includes country-level information about the number of network found with and without anti-spoofing. As mentioned in the introduction, recent studies showed a positive influence of disclosing social information on pro-social behaviour (Frey & Meier, 2004). Thus, we are interested in testing whether this relation holds also in the case of investments in cybersecurity, and especially if privately disclosing information about compliance (i.e. pinpointing that the majority of networks are already compliant) increases remediation. A field experiment will be designed to test the effectiveness of this type of private notification on the deployment of anti-spoofing filters.

2. *SQ2: What is the effect of publicly notifying non-compliant operators?*

Second, we investigate the effectiveness of public notification (i.e. publicly disclosing compliance information), and eventual differences between the private and public notifications. We suspect that public notifications may create additional incentives to remediate out of reputation concern and peer pressure. In order to produce such effect, we will publicly disclose information about compliance with anti-spoofing best practises, showing which networks are compliant and which are not. In particular, we will share this information with national CERTs[2], the network operator community and security bloggers, in order to evaluate if operators exposed to public notification react differently than those privately notified.

3. *SQ3: Can we identify characteristics of network operators that explain differences in remediation?*

Next, we seek to understand whether differences in remediation can be explained by organisational factors (e.g. the size of the network or the type of service provided) and socio-technical factors (e.g. the level of development of the ICT infrastructure or the presence and activity of cybersecurity institutions

[2] Computer Emergency Response Teams: group of experts who handle security incidents.

in a country). In fact, as the term *network operator* is used to indicate a variety of organisations, we investigate whether characteristics such as the country and the type of service provided, contribute to increase or decrease the effectiveness of the disclosure. To answer this question, a regression model is built on the data about remediation collected via the experiment.

4. *SQ4: What practical recommendations can be formulated on the base of the previous findings?*

Finally, we shall reflect on the previous findings in order to formulate concrete recommendations for defenders of Internet security, including researchers that everyday investigate new strategies to notify vulnerable parties, policy makers struggling to address the problem of IP spoofing and network operators.

1.3 Research approach

In order to address the questions just formulated, this research combines qualitative and quantitative research methods. We shall begin our investigation reviewing the literature on the economics of cybersecurity, a recent field of study that frames cybersecurity problems in light of economic concepts, with a particular focus on the role of actors' incentives. The combination of this literature and previous research on IP address spoofing will help us to identify the factors that contribute to operators' incentives to adopt anti-spoofing defences. On top of the literature review, these factors will be further investigated with interviews to network operators, in order to get a more practical insight on their perception of the problem.

Secondly, a field experiment is designed to test the effectiveness of the private and the public disclosure of information about compliance with anti-spoofing best practices as a way to notify non-compliant operators. In particular, we shall test two strategies to use such information. On the one hand, we seek to understand the effects of privately notify non-compliant operators providing compliance information (which operators are compliant, and which are not). On the other hand, we aim at publicly disclosing such information, testing if involving third parties may create additional pressure on non-compliant operators.

The experiment is based on the information collected by the Spoofer Project (CAIDA, 2018). Though this information is already publicly available on the Internet, we believe that better aggregating these results and using it as targeted feedback for operators might increase the chances of remediation. To disclose the information about compliance with anti-spoofing best practices, we design Infospoofing.com, a website on which we regularly release country-specific information about which operators were found with and without anti-spoofing filters. Observing operators' reaction to our website and tracking its visits is an important indicator for assessing the effectiveness of the notification, as they represent a proxy for *operators' intention to remediate*. In a way, we expect that operators visiting Infospoofing will be more likely to remediate.

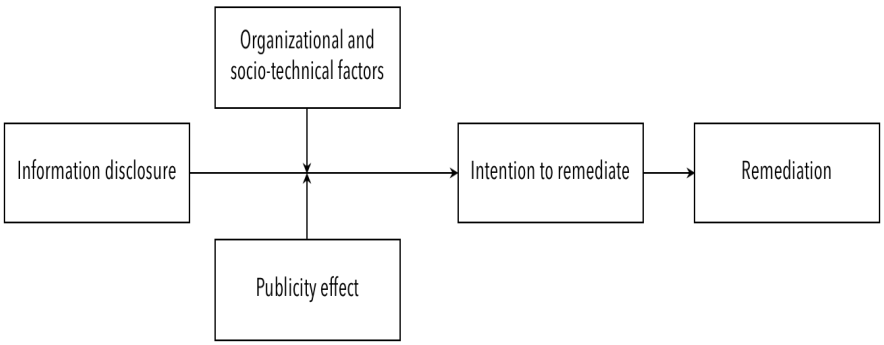
To test the effectiveness of our private and public notification, non-compliant operators are divided in three groups: a control group (to which no treatment is applied), one group for the private notifications

and one group for the public notification. We will send a mail notification to operators in both the private and public group, reporting the lack of SAV and providing information on the number of networks found compliant and not compliant. This mail notification also includes a link to our website, where more information is provided[3]. The difference between the two treatments is that for the public group we will share the website displaying information about non-compliant operators with additional third parties that might further instigate remediation. Thus, while differences in remediation between private group and control are due to the *information disclosure*, differences between the private and the public group are attributed solely to the *publicity effect*.

By looking at the remediation rate among the three groups, we seek to assess what is the effect of privately disclosing information, and if the public disclosure might further increase the chance of remediation. Furthermore, an explanatory analysis is conducted to investigate the role of *organisational and socio-technical factors* on the likelihood of remediation.

All in all, the approach taken in this research can be schematise in the conceptual model presented in Figure 1 (elaborated and further discussed in Section 3.3.2).

Figure 1. Conceptual model.



1.4 Contributions

1.4.1 Scientific relevance

This research contributes to the literature on the economics of cybersecurity, an interdisciplinary field of study that frames cybersecurity problems in light of economic theories. In fact, it has been argued that security failures occur as much for technological reasons as for *perverse incentives* (Anderson, 2001). Therefore, any analysis of security problems should start by studying actors' incentives (Moore, 2010). The economics of cybersecurity investigates the reasons that lead companies and users to under invest in security and is concerned with the design of interventions to stimulate additional incentives.

[3] Note that in the original design the notification also included a short questionnaire to investigate the reason of the lack of compliance and operators' intention to remediate. However, as we only received two responses, the questionnaire has been excluded from the research.

In particular, it has been acknowledged that gathering, analysing and sharing security information play a key role to improve Internet security (Gordon, Loeb & Lucyshyn, 2003; Gal-Or, & Ghose, 2005). Some authors have further suggested that publicly releasing information about organisations' security performances might create additional incentives out of reputation effect (Tang, Linden, Quarterman & Whinston, 2013). Our research fits in this debate and attempts to shed light on the effectiveness of disclosure policies as a mean of instigating security behaviours.
Notably, it might be the case that the vulnerability information itself is already available, and what is missing is a proper aggregation and feedback system (Tang et al., 2013). In our study, we use vulnerability data collected and already publicly available on the Internet. This type of approach to vulnerability notification can apply to a variety of security issues, for which information can be collected, for example, via public policies on mandatory disclosure.

In practice, our goal is to increase the adoption of anti-spoofing measure, thus contributing to reduce Internet's susceptibility to attacks based on IP spoofing. In this regard, every instance of remediation represents a small step towards a more secure Internet ecosystem. Moreover, remediation aside, public disclosure tends to increase transparency and awareness about the problem, which may further attract the attention of operators, researchers and regulators.

1.4.2 Deliverables

The final outcome of this research will be a report composed of 6 chapters, and the structure of the research is shown in Figure 2.

This research proposal will represent the first chapter of the final report.

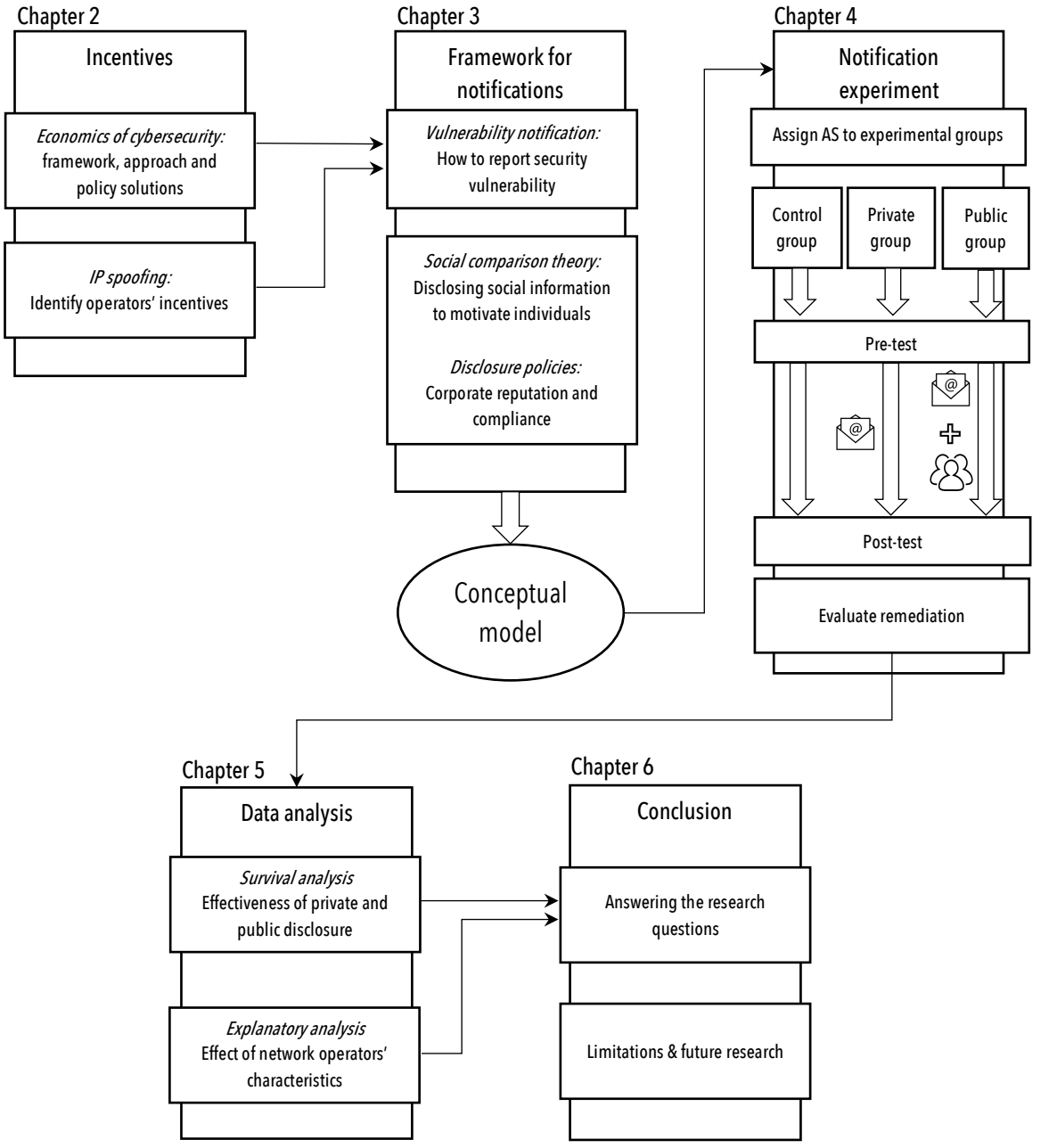
In the second chapter, we shall provide a review of the literature on the economics of cybersecurity, as well as a discussion on the problem of IP spoofing. In particular, we will discuss how economic barriers such as externalities, information asymmetries and misaligned incentives affect the diffusion of anti-spoofing mechanisms.

Chapter three will discuss the importance of sharing security information and will focus on the practice of vulnerability notifications. In addition, we will explore the use of reputation and peer pressure to increase incentives to comply with norms and regulations. At the end of the chapter, the research questions and the conceptual model of this research are developed, and the empirical hypothesis are formulated.

The fourth chapter will discuss the research methodology. First, we will introduce the source of data and the metrics used to quantify and aggregate the information we want to release. Secondly, the design of the notification experiment is elaborated.

In chapter five, the results of the experiments are analysed and presented. Finally, chapter six concludes the research by discussing the policy implications of our findings, pointing to the limitation of our work and suggesting future direction for further research.

Figure 2. Research structure.



Chapter 2:
An economical insight on IP address spoofing

In this very moment, somewhere on the Internet, a network is under attack. The rate, the variety and the impact of cyberattacks has dramatically increased over the years, and today cybercrime is estimated to cost the world's economy $600 billion, an increase of almost 35% from 2014 (McAfee, 2018). If, on the one hand, attackers are improving day after day, rapidly adopting new technologies and exploiting the growing number of targets connected to the Internet, on the other hand defenders and law enforcement are struggling to keep up, making cybercrime an attractive low-risk and high-revenue option.

To better understand the dynamics that lead security systems to fail so often, scholars have started looking to security problems through the lens of economic theories, focusing on the structure of incentives that drive actors' decisions in the cyberspace. In particular, a key observation is that security failures are likely to occur when the party responsible for the security of a system does not suffer the consequences of its failure (Anderson & Moore, 2006). For instance, while individual users may be inclined to invest on anti-virus programs to protect their devices, it is unlikely they will spend even little amounts for software that prevent their devices from being used to attack other parties (Varian, 2000). Likewise, companies are unlikely to invest in security measures that do not present an adequate return on investment, despite their adoption might be beneficial for the cyberspace as a whole. As a result, the diffusion of important security standards, protocols and best practices might take a very long time (see, for example, the case of DNSSEC and S-BGP, two security upgrades for core Internet protocols whose adoption is still partial (Anderson & Moore, 2006)).

The focus of this thesis is on the case of IP source address spoofing, a fundamental problem in the Internet's architecture. Despite being a well-known issue for more than 30 years (Morris, 1985), IP spoofing remains a popular attack method, due to a misalignment in the incentives of the actors involved in the remediation of the problem (Beverly et al., 2013).
Before discussing the problem related to IP spoofing in details, its consequences and mitigation, the first part of this chapter introduces the field of the economics of cybersecurity, presenting an overview of the factors that make cybersecurity problems so hard to solve (Section 2.1). These concepts will be used in the second half of the chapter to discuss the slow adoption of anti-spoofing measures from an economical perspective (Section 2.2).

2.1 The Economics of Cybersecurity

The security landscape we observe today is the result of decades of incremental and decentralised decisions made by actors in the cyberspace. The variety of actors and the complicated patterns of their

dependencies have made the cyberspace a highly complex system, often described in terms of an ecosystem to highlight these interdependencies and the co-evolution of actors (van Eeten & Bauer, 2008). Actors in the Internet ecosystem include network operators, hardware manufacturers, software developers, application and service providers, security providers, (inter)national governance agencies and various type of users (van Eeten & Bauer, 2008). The resulting set of actors is fairly heterogeneous with regard to both their skills and to the motives that drive their actions.

Every day, each actor takes security related decisions. Given the interconnectedness of the cyberspace, their choices end up affecting the whole Internet ecosystem (van Eeten & Bauer, 2008). Although the decisions of each actors may be rational considering his individual costs and benefits, the outcome of these decisions might be far from desirable when considering costs and benefits aggregated to the level of the cyberspace. And as the overall level of security is the results of decentralised decisions taken by individual actors, it is crucial to ensure that each actor has a well-defined structure of incentives to drive his decisions towards outcomes that are desirable for the whole cyberspace.

Born in the early 2000, the field of the Economics of Information Security puts under the microscope the factors that influence actors' security decision making (i.e. their *incentives*[4]). Incentives are embedded in the economic fabric that ties actors in the cyberspace together, including the economic condition of the market, the interrelationships among actors, the legal framework and the set of social norms (van Eeten & Bauer, 2008). In 2001, Ross Anderson argued that, in spite of the relevance of technological factors, security failures occur as much because of *"perverse incentives"*, claiming that: *"Many of the problems can be explained more clearly and convincingly using the language of microeconomics: network externalities, asymmetric information, moral hazard, adverse selection, liability dumping and the tragedy of the commons"* (Anderson, 2001: 358).

In the next sessions, three major economic barriers to cybersecurity are discussed. Specifically, Section 2.1.1 deals with externalities in cybersecurity, Section 2.1.2 explains how information asymmetries might lead to under investments in security, and Section 2.1.3 describes how misaligned incentives might arise. Next, in Section 2.1.4, we briefly review the policy options that have been proposed to cope with these economic barriers to cybersecurity.

2.1.1 Externalities

In economic theory, *externality* refers to any cost (negative externality) or benefit (positive externality) perceived by a party who has no control over the factors that create it. A classic example of negative externality is air pollution, where the costs of polluted air encumber on the whole society, and not on the party accountable for the pollution. The presence of externalities is typically considered as a type of market failure, especially in combination with public goods[5] (e.g. fresh air) (Buchanan & Stubblebine, 1962; Camp, & Wolfram, 2000).

[4] From an economic standpoint, the behaviour of people is influenced by incentives: financial and non-financial factors that encourage to prefer one choice over its alternatives. Incentives are the expectations that motivate certain behaviour or courses of action above others.

[5] A public good is at the same time a "non-rivalrous" good (i.e. the individual consumption does not affect its availability for other individuals) and "non-excludable" good (i.e. no individual can be excluded from using it).

Arguably, cybersecurity can be considered a quasi-public good[6]. In fact, despite it is largely provided by private actors, it also presents strong public goods features (van Eeten & Bauer, 2008). In economics, the provision of public goods represents a challenge due to the *free rider problem*: individuals may be able to enjoy benefits of a public good without participating to its costs, resulting in an under-provision of the good itself (Anderson & Moore, 2007).
In the case of cybersecurity, security investments of individuals may contribute to increase the overall level of security of the cyberspace, generating positive externalities for other parties which, in turn, might discourage additional investments (Kunreuther & Heal, 2003). Hence, free-riding behaviours are likely to occur, especially when security depend on the party with the highest benefit-cost ratio (Varian, 2004).

In practice, there are many instances of externalities related to insecurity. A particularly good example is the case of *botnets*, networks of compromised devices under the control of an attacker. Devices in a botnet are typically used for a variety of fraudulent schemes, from spreading spam and phishing contents to launch denial of service attacks, often without awareness of the owners of the compromised devices. The problem with botnets is that the target of the criminal activity is not the owners of the infected devices, but someone else. This originates negative externalities towards the victim: owners of compromised devices have no reason to improve their security as long as somebody else is attacked (Moore, 2010).

Sometimes externalities manifest themselves in other market mechanisms. For instance, the market for software presents similar features to the market of many information goods, which might be affected by direct or indirect *network effect* (Shapiro & Varian, 1998). Direct network effect is a condition for which the value of a product increases as its user base gets larger (e.g. the utility of having a telephone grows with the number people owning one). Indirect network effect, instead, refers to an increase in the value of a product because, as the user base grows, more complementary products are available (e.g. the more people own a smartphone, the more mobile applications will be developed, increasing the utility of owning a smartphone).
In the opening of this chapter, we have mentioned DNSSEC and S-BGP as examples of security protocols whose widespread adoption have taken a very long time. In fact, these protocols do not yield significant benefits until a critical mass of users has adopted them, which is an example of network effect. As a result, there was little incentives to be among the early adopters (Moore, 2010).

The presence of strong network effects can have severe implications for security, due to a condition known as *customer lock-in* (Gottinger 2003) (Lookabaugh & Sicker 2004). Lock-in happens when customers are dependent on a product because of high switching costs: organisations using a particular software might face costs for switching to another software (i.e. re-installation, training) that are so high to discourage the change. Vendors enjoy lock-in situations, since they represent a barrier for new firms to enter the market. Therefore, vendors will try to reduce the time-to-market of their product, in order to build a solid position on the market and to create customer lock-in. In doing so, often security requirements are ignored in the early stage of product development (Anderson & Moore, 2006).

[6] For a detailed dissertation of cybersecurity as a public good, see (Powell, 2005) (Mulligan & Schneider, 2011) (Asllani, White & Ettkin, 2013).

2.1.2 Asymmetric information

Incomplete and asymmetric information represents another factor that contribute to the failure of the market for security. Many economic models are based on the assumption of *complete information*, a condition in which all market players have the same, complete knowledge about other actors. When this assumption does not hold, and information is not uniformly distributed among actors, market failures are likely to occur, as famously shown by George Akerlof with his *market for lemons*[7] (1978).
Asymmetric information affects cybersecurity to the extent that consumers cannot determine the security of a product or service. It has been pointed out, for example, that the market for secure software (e.g. anti-virus) resembles a market for lemons, since vendors know the level of security of the software, while consumers cannot assess it. As consumers refuse to pay a premium for better products, vendors are discouraged from offering them, driving down investments for more secure products (Anderson, 2001). A similar logic applies also to the case of consumers looking to purchase Internet services, such as Internet access.

Situations like the market for lemons are not the only way in which asymmetric information affects security decisions. In general, incomplete or unreliable information leads to a poor allocation of resources. In fact, some organisations are likely to underreport security breaches for fear of losing customers, others do not cooperate with law enforcement on cyber espionage incidents since their reputation or share price may drop. This combination of unreliable information and *FUD* (fear, uncertainty, doubt) can have dramatic impacts on security investments. Without a clear comprehension of the cyber *insecurity* problem, of its magnitude and main threats, palliative solutions will be sold to ill-informed buyers, while defences against the most substantial threats will not be developed (Moore, 2010).

2.1.3 Misaligned incentives

As mentioned in the introduction to this chapter, security failures are likely to occur every time that the party responsible for the security of a system does not suffer when the system fails.
In 1993, Ross Anderson observed that patterns of frauds in ATMs depended on the party liable for them. When disputing a transaction with a customer, banks in the US had the obligation either to prove that the customer was wrong or to offer him a refund. Conversely, banks in the UK had no such obligation. Instead, it was the customer that had to prove that the bank did a mistake. Given that proving the wrongdoing of the bank was almost impossible for customers, banks in the UK became careless to customers' complaints. As a result, an avalanche of fraud flooded banks in the UK, who spent much more on security than their American counterparts (Anderson, 1993). Economically speaking, this is a case of *moral hazard* (i.e. a party takes additional risk when he is not subject to the negative consequences), which is the symptom of an ill-allocation of risk or liability (Anderson, 2001).

[7] Imagine a market for used cars, composed by an equal number of good cars (each worthy €1000) and bad cars (each worthy €2000). If buyers cannot distinguish between good and bad cars (a.k.a. *lemons*), they are going to offer an initial price of €1500, which averages the value of good and bad cars. However, as no seller would sell a good car to this initial price, owners of good cars will leave the market, resulting in a drop of the market price for cars until, eventually, the market will consist of only lemons sold to €1000. The key takeaway is that consumers are unwilling to pay an extra price for quality they cannot measure, resulting in a market of low-quality products (Anderson & Moore, 2007).

Another case of incentive failure refers to the so-called *tragedy of the commons*[8], an economic problem in which decisions pursuing actors' self-interest get outcomes that are contrary to the common good. It has been argued that distributed denial of service attacks can be associated with the tragedy of the commons (Anderson, 2001). In this type of attack, a large number of compromised devices redirect their traffic to the victim, with the aim of overloading his system until he is cut-off of Internet connection. While users buy antivirus to protect themselves, unlikely they will spend to prevent attack scenarios in which they are not the victim.

In general, misaligned incentives might arise every time security is traded-off. Previously, we have mentioned the trade-off between security and time-to-market in the software industry, but there are plenty of similar cases. Consider, for instance, the trade-off between security and efficiency in online banking. On the one hand, consumers enjoy the efficiency of online banking. On the other hand, online banking exposes them to a higher risk of fraud. Therefore, as security comes to cost, it appears rational for individuals and society to tolerate up to an optimal level of insecurity (i.e. until any drop in risk of additional security measure does not cancels out the advantages of efficient operations). In fact, albeit risks of online fraud could be eliminated by simply not using online banking, society would actually be worse off, as the total cost of offline banking would counterbalance the losses related to fraud (Moore, 2010).
However, misaligned incentives arise when the actor making the trade-off does not suffer the consequences of the attack, leading to suboptimal choices of where the trade-off is made. Unluckily, this is the case for many security decisions.

2.1.4 Proposed policy solutions

In the previous sections, we have discussed three main economic barriers to cybersecurity, namely externalities, asymmetric information and misaligned incentives, showing potential failures associated with each of them. Though we discussed each barrier separately, in many security problems they occur together, often with a combined effect. To cope with these economical shortcoming, some sort of intervention appears to be necessary. A traditional solution is governmental intervention, which, nevertheless, also presents some limitations. Since the Internet spans over the boundaries of national jurisdiction, its governance requires a high degree of cooperation. On top of that, governmental intentions might be biased as well (see, for example, the surveillance/privacy dilemma). Over the years, other types of policy options have been designed to fix cybersecurity problem. In this section we review the most relevant ones.

Regulations and self-regulations

Regulatory institutions, like governmental agencies and telecommunication authorities, can play an important role in shaping the cybersecurity posture of a country. For example, since the mid 1980s,

[8] The term "Tragedy of the Commons" appeared in a 1968 paper, which discusses how each individual tries to reap the largest benefit from a shared resource, creating an economic situation in which the action of one party generate a substantial benefit for that party, and a small cost for everyone else. Imagine a group of herdsmen who feed their cattle on a common field. The tragedy happens when a herdsman adds a new cow to his herd: he will get the benefit of having an additional cow, but all the cows will now have less to eat (Hardin, 1968).

certification schemes have been proposed to tackle the information asymmetries in the software market (first, the US with the "Orange Book", and then NATO with the "Common Criteria"). However, both attempts had a short life, most notably due to *adverse selection* (Anderson & Moore 2006). In addition, Ghose and Rajan (2006) showed that heavy-handed regulation, i.e. make compulsory investment in security compliance, can have strong and unintended consequences, including distorting security markets and decrease competition.

Traditionally, the fight against cybercrime has been organised in terms of voluntary and self-regulatory actions (van Eeten & Bauer, 2008). Voluntary action can take many forms, from reporting software bugs to maintain blacklist[9] (Jhaveri, Cetin, Gañán, Moore, & van Eeten, 2017). Self-regulation is another approach that has been tried, often with different outcomes (i.e. it has proved effective in terms of patch-management, but it has also failed in the case of website approval seal) (Anderson & Moore, 2006).
A particularly interesting form of self-regulations are social contracts and collective actions. Since actors in the Internet ecosystem are in a prisoner's *dilemma situation* (i.e. everyone is worse off if they take individually rational decisions), cooperation and coordination among actors is critical (Moore & Clayton, 2008). One way to put the good of the commons in front of personal gains, and therefore to avoid Hardin's tragedy of the commons, is to have actors participating in a social contract: actors take a commitment to the public good and expect the same from others (Internet Security Alliance, 2008).

Assigning liability

In the previous section we have shown that misaligned incentives between who is responsible for security and who enjoys its benefits (or suffer the costs of its lack) is a major cause of failure for security systems (Anderson & Moore, 2006). One possible way to address misaligned incentives is to establish a liability regime in order to bring back the responsibility of security on the party that can best manages the risk. For example, we have seen how unsecure software might be brought on the market as a result of vendors' battle for dominance. In this case, it might seem rational to assign liability on the software vendors. So, should Microsoft be considered accountable for the consequences of attacks exploiting vulnerabilities in Windows operating systems? One's first inclination might be to argue that the threat of monetary damage brought about by legal actions would create a very powerful incentive for Microsoft to secure their products. Nevertheless, such blanket assignment of liability could generate significant side effect. First and foremost: *"If each new line of code creates a new exposure to a lawsuit, it is inevitable that fewer lines of code will be written"* (Moore, 2010: 108), slowing down the pace of innovation. In addition, this type of software liability will also bring negative consequences for the part of the industry that develops free and open software, representing a disincentive for users to contribute. Finally, software is inherently *buggy*, and it would be unrealistic to think that developers could eliminate all the vulnerabilities overnight.

Indirect and intermediary liability

Sometimes, third parties can be held responsible for the actions of other actors. For example, employers can be considered responsible for the wrongdoing of their employees. In cybersecurity, assigning liability

[9] lists of compromised IP addresses or phishing URLs

to attackers can be difficult because they may operate anonymously beyond the reach of the law, and, even if caught, they cannot afford to pay for the damages they cause. Thus, intermediary liability appears a viable solution, especially if there are third parties in the Internet ecosystem in the position to detect and prevent exploits. Ultimately, by internalising the costs of insecurity, these third parties would reduce the impact of negative externalities (Lichtman & Posner, 2006).
Some authors have argued that Internet Service Providers (ISPs) are in a good position to act as intermediary. For example, van Eeten and colleagues analysed the role of ISPs as intermediary in botnet mitigation (van Eeten, Bauer, Asghari, Tabatabaie & Rand, 2010). Varian (2000) proposed that ISP should be bear the costs of DDoS attacks originating in their network. Lichtman and Posner (2006) argued that ISPs should be considered liable (at least partially), because they control the gateway through which malware is diffused. Nevertheless, it is important to notice that ISPs are already involved in the mitigation of several Internet problems, from phishing and spam to major botnet takedown, from blocking offensive material to infrastructure resilience. Despite they clearly have a central role in Internet security, ISPs cannot be responsible for the security of the whole ecosystem.

Cyber insurance

Insurances are a traditional solution to problems of risk management. A robust system of cyber insurance could bring multiple benefits to the Internet ecosystem (Bohme & Schawrz, 2010; Schneier, 2004). First, insurances incentivise individuals and organisation to adopt proper security measures, since risky behaviours are discouraged by higher premiums. Secondly, insurances have an incentive to collect data on security incidents during claims and disputes, which would address the problem of unreliable and asymmetric information.
In spite of these benefits and an initial optimism for cyber insurance, they never really took off (Bohme & Schawrz, 2010; Moore 2010). In fact, the presence of information asymmetry makes very hard for insurers to quantify cyber risk (Shetty, Schwartz, Felegyhazi, & Walrand, 2010). In addition, the design of insurance policies in the cyberspace is complicated by the interdependencies of actors, which make cyber risk a *hyper-risk* that might result in cascading failures (Helbing, 2013).

Security breach disclosure

As incomplete and unreliable information represents a major barrier to cyber security, policies aimed at disclosing and sharing information may represent a valuable tool to break down this barrier. In recent years, many US states and a number of countries have made the disclosure of security breach mandatory (Tang et al., 2013). Besides being motivated by community's *"right to know"* (Moore, 2010: 108), public notification of security breach represents a powerful incentive for organisations to adopt additional security measure (e.g. new access controls, auditing measures, and encryption (Mulligan, 2007)). Campbell, Gordon, Loeb & Zhou (2003) found that the disclosure of security breaches leads to a loss of market values only if the breach involved unauthorized access to confidential information. Also Acquisti, Friedman & Telang (2006) observed a drop in the stock price of a firm after a breach was reported. In a study considering time differences in the enactment of mandatory disclosure laws in US states, Romanosky, Telang & Acquisti (2011) found a small, yet statistically significant decrease in fraud rate after the adoption of such norms.

Vulnerability disclosure

Another interesting debate refers to whether newly discovered security vulnerability (and vulnerability information in general) should be publicly disclosed. While supporters of this idea argue that mandatory disclosure encourage vendors to release patches, opponents of vulnerability disclosure argue that it might provide attackers with information they may not obtain otherwise. However, there are some empirical evidences in support of the disclosure of vulnerability. In particular, Arora, Krishnan, Telang & Yang (2010) found evidences that the public disclosure can accelerate patch release. Moore & Clayton (2011) showed that phishing website reported on a public blacklist are less likely to be compromised when compared to website only known in small communities. Some have also taken a step further: Tang et al. (2013) publicly disclosed data on organisations' outgoing spam in form of peer ranking, in order to increase organisations' incentives to clean up out of reputation concern.

More generally, disclosure policies aimed at publicly revealing organisations' wrongdoing (aka *naming and shaming*) have been proposed as a means to prompt compliance with norms. Such policies are considered as an alternative type of regulatory intervention, along with "command and control" and "market regulations" (Florini, 2008; Gupta, 2008). Van Erp (2011) argues that, beside increasing transparency and alleviating information asymmetries, disclosure policies are effective for three reasons: first, they impose a reputation sanction to offenders; second, they work as a deterrent for other organisations; and third, they help setting the boundary between acceptable and not acceptable behaviours. We will devote more attention to disclosure policies in the next chapter (Section 3.3).

So far, we have discussed how economic barriers like externalities, information asymmetry and misaligned incentives may complicate cybersecurity problems. Then, we have briefly reviewed the main policy interventions that have been proposed to overcome such barriers. Unfortunately, accurately discussing each intervention would require much more space than what we dedicated. However, for each intervention, we tried to summarize potential benefits and shortcomings. Our aim, in this first part of the chapter, was to show that, contrary to what many people can think, cybersecurity is not only a problem of compliance and technology, but actors' incentives need addressing. Without proper incentives to do security, rational individual choices can have disruptive outcomes on the Internet ecosystem.
In the next section we will focus on the problem of IP spoofing, which represent an interesting case for seeing many of the dynamics previously discussed in action.

2.2 Economics of IP spoofing

In this section we dive into IP spoofing. First, the technical side of the problem is explored. Section 2.2.1 introduces what is IP spoofing, Section 2.2.2 illustrates a typical attack scenario, Section 2.2.3 discusses the possible mitigations technologies and Section 2.2.4 describes how to measure deployment of anti-spoofing filters. Secondly, we study the incentives side of the problem, focusing on network operators, the central actors in the mitigation of IP spoofing. Section 2.2.4 provides an overview on the factors that generally contribute to network operators' incentive. Finally, Section 2.2.5 investigates the

incentives of operators to deploy anti-spoofing and explore possible intervention to increase such incentives.

2.2.1 IP source address spoofing

At first glance, the Internet might seem one global, homogeneous network. In reality, it is an assortment of several independent networks, connected each other with an elaborate system of routing and addressing. Information on the Internet flows across different networks thanks to routers, devices that steer information packets to their destination. A set of routers under a single technical administration forms an *autonomous system*[10] (AS), the fundamental building block of the Internet.

Beside its data content, Internet packets include metadata about the packet itself, like the source and the destination IP address. While the destination address indicates the machine to which the packet must be delivered, the source address is used to identify the sender of the packet, so that the recipient can respond to the right machine.
By their design, routers, in order to deliver a packet, generally examine only the field containing the destination address. Arguably, this is the most fundamental vulnerability in the Internet architecture, as it does not include any notion of *packet-level authenticity* (i.e. the identity of the sender is not verified) (Beverly et al., 2009).

IP source address spoofing, or IP spoofing in short, is the practice of creating Internet packets with a fake source address, with the aim of hiding sender's identity or impersonating another computer system (Internet Society, 2015).
Applications with sufficient privileges can easily create packets with a spoofed address, and often these packets can travel across the Internet and reach their destination. Obviously, the use of a fake source address hampers regular communication: once the spoofed packet is received, responses from the destination will be sent to the (fake) source address indicated in the header of the packet, not to the actual sender. Despite the lack of practical uses for normal communication operations, attackers can abuse IP spoofing to launch a variety of attacks, chief among which are reflection and amplification DDoS attacks.

2.2.2 Reflection and amplification DDoS attack

Reflection and amplification DDoS attacks represent a major plague for the cyberspace. Since the dawn of the Internet they have increased in size and frequency, so much so that today it is almost difficult keep track of the new records. To give an idea of the exponential grow of such attack, in 2014 Cloudfare reported a *massive* attack of over 400 Gbps (Prince, 2014). Two years after, in 2016, Brian Krebs was victim of a 620 Gbps DDoS attack (Krebs, 2016). As for 2018, in February GitHub was hit by a DDoS of 1.35 Tbps (Kottler, 2018), setting a new record that lasted until, just 5 days later, Arbor Network confirmed a 1.7 Tbps DDoS attack (Morales, 2018). All the attacks mentioned are reflection and

[10] *"An AS is a connected group of one or more IP prefixes run by one or more network operators which has a SINGLE and CLEARLY DEFINED routing policy"* (Hawkinson, Bates, 1996: 3).

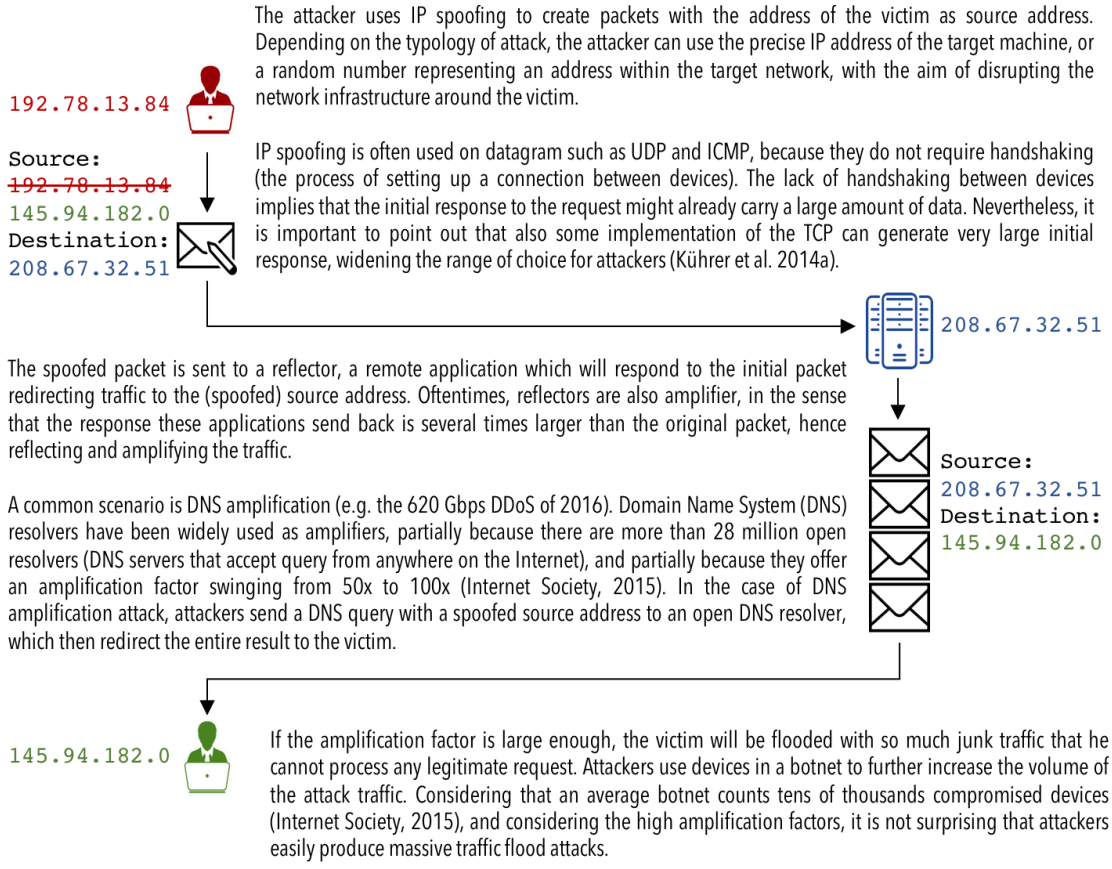
amplification DDoS and are enabled by IP spoofing. In this section, the typical structure of such attacks is explained.

Anatomy of an attack

A denial-of-service (DoS) attack is a type of cyber-attack in which the goal of the attacker is to cut off Internet service to the victim. DoS attacks are typically accomplished by flooding the victim with a very large amount of traffic, consuming his bandwidth until the victim has no resources left to process regular traffic. In a more sophisticated attack scenario, attackers can launch a distributed denial-of-service (DDoS) attack by flooding the victim with traffic originating from multiple source (e.g. using devices in a botnet), making it infeasible to block the attack by simply blocking one source (Kührer, Hupperich, Rossow, & Holz, 2014a).

As mentioned, a particularly damaging category of DDoS are *reflection and amplification DDoS attack*. In this type of DDoS attacks, the attacker spoofs the IP address of the victim, creating a packet using the address of the victim as source address. The spoofed packet is then sent to an intermediary application which, in turn, will send back its response to the (spoofed) source address in the initial packet. Specifically, a typical reflection and amplification DDoS attack consists of the three steps shown in Figure 3.

Figure 3. Schematization of a reflection and amplification DDoS attack



It is important to point out that the use of IP spoofing lies in the "pre-phase" of the attack: the spoofed request is the initiator of the attack. In fact, the attack traffic generated by the reflector is oftentimes legitimate traffic. Surely, a very large amount of traffic, but legitimate (e.g. the response of a DNS). Thus, the only way to fully prevent DDoS is by stopping spoofed packets from leaving the source network.

Disrupting the attack chain

The mitigation of reflection and amplification DDoS attack can be approached by multiple points of view: preventing spoofing and closing amplifiers.

First and foremost, network operators can block the initial spoofed packet before it reaches the amplifier. Preventing attackers from creating packets with a spoofed source is technically unfeasible (i.e. attackers can always create packets with a fake source address). However, the operator of the network in which the attack is launched can drop the packets that have a source address which does not belong to that network, preventing their ingress in the public Internet. This type of filtering is known as Source Address Validation (SAV), and will be further discussed in the next subsection.
It is important to notice that IP spoofing (or, to be more precise, the lack of anti-spoofing filters) might not entirely fit the definition of security vulnerability. In fact, a *spoofable* network is not per se vulnerable to attacks[11], but it can be exploited during an attack against other networks. Therefore, also the term *remediation*, typically associated with the remediation of a vulnerability, might be used improperly in this context. Nevertheless, as IP spoofing has been defined as *"a critical problem in Internet architecture"*, we will use the term *remediation* to imply the deployment of anti-spoofing filters, in the sense of the remediation of the problem.

A second solution would be to focus on the reflector/amplifier, by putting some restrictions on the clients that can send requests. For example, in the case of DNS amplification, attackers use open DNS resolvers, which are publicly accessible resolvers that accepts query from anyone on the Internet. However, since the majority of legitimate DNS queries comes from clients in the same network of the resolver, implementing an access control list on the DNS might "close" an otherwise open resolver (Internet Society, 2015). Another solution would be implementing Response Rate Limiting (RRL), in order to lower the rate of response to possible malicious queries (Internet Society, 2015).

Nevertheless, working on the reflector side is difficult for a variety of reasons. First, considering the case of DNS amplification, closing all the open resolvers can be a very long process, albeit theoretically desirable. Moreover, besides UDP, in 2014 14 additional protocols were identified to be susceptible to bandwidth amplification, some with an amplification factor of 4670x (Rossow, 2014). On top of that, the most recent records of amplification DDoS attack was achieved using Memcached (running on UDP), with an amplification factor of more than 51,000x (Morales, 2018). Through the use of UDP on Memcached was immediately disabled, imagining to be able to remove all these sources of amplification would be unrealistic.

[11] Naturally, there might be attack scenarios in which the network lacking anti-spoofing is itself the victim of the attack (e.g. spoofed access to network equipment).

2.2.3 Ingress filtering and source address validation

The solution to prevent spoofed packets to leave the source network is to do ingress filtering: network operator should validate the source IP address before forwarding the packets to the Internet (a type of filtering known as *source address validation*, schematised in Figure 4). In this section we review different ways to implement source address validation.

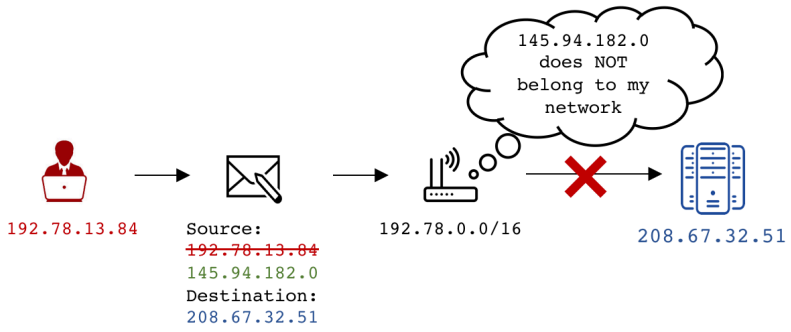
BCP38

Network operators must implement ingress filtering according to BCP38 (RFC 2827). BCP38 is a Best Current Practice aimed at blocking spoofed packets at edge of the source network, before they enter the public Internet. In fact, only the routers in the network of the attacker know which IP addresses belong to that network (and thus which address to expect as legitimate source address). Therefore, those routers near to the real source can drop the packets with a spoofed address. This is typically achieved by implementing an ingress access list containing the IP prefixes that are acceptable on a given router interface. In particular, BCP38 should be implemented on the routers the edge of a network, or better, on every edge of every network. Such type of filtering is also known as Source Address Validation (SAV), and has been proposed in 1998 (Ferguson & Senie, 2000).

BCP84

However, there are some scenarios in which BCP38 might create problems, for example in the case multihomed networks (i.e. networks connected to more than one network interface). To address these cases, BCP38 has been updated by BCP84 (RFC 3704), which introduces Unicast Reverse Path Forwarding (uRPF) as a mechanism to simplify the implementation of dynamic ingress access lists. uRPF uses additional routing information provided by routers adjacent to the one on which it is implemented (namely RIB, routing information base, and FIB, forwarding information base) to further restrict the set of acceptable sources on a given network interface (Baker & Savola, 2004). Still, also uRPF presents some complications, as it might significantly reduce performances of the network. Moreover, since its strength lies in the information provided by nearby routers, uRPF becomes less effective and more fragile the further it is implemented from the source of the spoofed packet (Internet Society, 2015).

Figure 4. Schematisation of source address validation



Source address validation improvement

Both BCP38 and BCP84 implement source address validation at a IP network layer, both operate with an address range on router interfaces and have to be implemented at the edge of the network. However, preventing IP spoofing more locally and closer to the hosts would be more robust and would also limit the scope of possible attacks (McPherson, Baker & Halpern, 2013). For this reason, in 2013, RFC 6959 proposed Source Address Validation Improvement (SAVI), an initial approach to implement anti-spoofing in a local network segment (operating with IP addresses and MAC addresses on switch ports).

2.2.4 Measuring deployment of source address validation

In general, there is plenty of information about DDoS activity, in terms of statistical data, reports on attacks and information about amplifier. Nevertheless, these reports do not often go beyond the reflectors that generate the attack traffic. Very little information is available on the actual source of the attack: the network that enable the whole attack chain by not deploying adequate anti-spoofing measures (Internet Society, 2015).

There are different methods to measure if a network deploys ingress filtering. The first relies on an insider running a testing application (Beverly & Bauer, 2005). Secondly, traceroute data can sometimes be used to determine the lack of SAV (Beverly & Bauer, 2007). Recent research showed that traceroute data can be used to reveal the absence of filtering by looking for routing loops (Lone et al, 2018). Finally, there is the Open Resolver Project, which uses a specific DNS implementation quirk typical in some customer-premises equipment to show lack of ingress filtering (Open Resolver Project, 2013). The rest of this section is dedicated to describing the first methodology, running a test application from inside the network, because it represents be the source of data that will be used in the rest of this research.

The Spoofer Project

The main technique to infer the presence of ingress filtering on a network is to run a test application that attempts to sends packets with a spoofed source address to a central server. Researchers of the Center for Applied Internet Data Anaysis (CAIDA) developed such test application in 2005 and have promoted its use ever since with the aim of understanding the extent of deployment of source address validation.

They have produced a large set of measurement and display all the test results from mid 2016 aggregated on different level (IP prefix, autonomous system and country), though the IP addresses that did the test are anonymised for security concerns. Example of test results are presented in Appendix 1.

Once installed on a host with sufficient permissions, the application tries to send a series of UDP packets with a spoofed source address. These packets are addressed to a server managed by CAIDA, where they are recorded together with other metadata on the test for later retrieval. In essence, if the spoofed packets arrive to the destination, it means that the source network does not properly implement anti-spoofing filtering. Importantly, once installed on a device, the application automatically runs both in the background once a week, and every time the device is connected to a new network. In this way, the state of a network is monitored over time.

After sending the test UDP packets with a spoofed source address, the application establishes a TCP connection with the server, in order to exchange the test result and conclude the test.

The test is performed crafting multiple spoofed packets, each using a type different source IP addresses specifically chosen in order to test common filtering policies. An overview of the IP addresses used by the application is presented in Figure 5.

Firstly, the test is run using an unallocated IP address. Unallocated addresses are addresses which have not been assigned to any organisation yet, and that therefore should not appear in any routing table. Secondly, the application creates a spoofed packet using a private IP address. Private addresses are legitimate addresses that must be used in private networks such as Local Area Networks (LANs), as prescribed by RFC 1918 and RFC 4143. Then, a valid and allocated address is used as source address for the spoofed packet. Differently from the previous cases, this time the address appears in the global routing table, albeit assigned to another organisation. Lastly, the application tries to measure the granularity of any applied filter, by spoofing IP addresses from netblocks incrementally adjacent to the one on which it is installed. This procedure starts by spoofing an adjacent /31 (i.e. the host's address +1), until reaching an adjacent /8 (Beverly & Bauer, 2005).

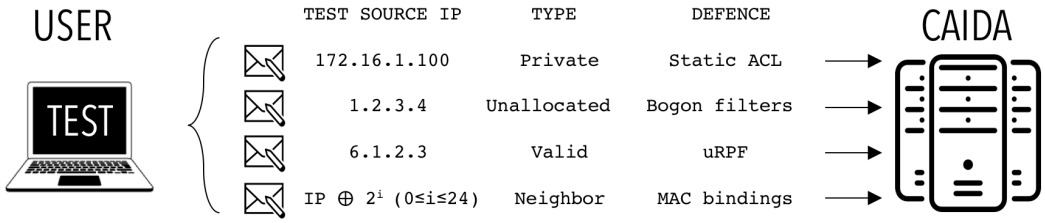
The test has four possible outcomes:

- *received*: the spoofed packet was received, which means that source network does not implement ingress filtering;
- *rewritten*: the spoofed packet was received but the original source address was changed en-route, which indicates the presence of a Network Address Translation[12] (NAT) that rewrites the header of the spoofed packet;
- *blocked*: the spoofed packet was not received, but the TCP packet (unspoofed) was, meaning that the spoofed packet was dropped by an in-network filter;
- *unknown*: neither spoofed nor unspoofed packet was received.

It is worthy to briefly discuss what happens when the test is run from a client behind a NAT. In this case, the results vary depending on the type of NAT implementation (Beverly et al., 2009). In general, if an intermediary NATting device rewrites the source address of the spoofed packet, it is not possible to infer the presence of anti-spoofing filters. In case of reflection attacks, the spoofed packet reaches the reflector, which response is addressed to the NATing device, thus preventing the success of the attack. However, there is no standard for a NAT to deal with spoofed packets, and different implementations may behave differently. In some cases, the spoofed packet is dropped, either by the NAT itself or by an in-network anti-spoofing filter. In other cases, the spoofed packet arrives to destination with the original source address unchanged, despite the NAT.

[12] NAT is the practice of remapping IP addresses from a network domain to another. Some networks internally use IP addresses that cannot be used outside that network, either because these internal addresses are not valid outside or because the internal addressing must be kept private. In a typical home network, for example, private addresses are assigned by a home router to each device in that network. When a device wants to communicate with Internet outside the home network, it sends a packet (with its private source address) to the home router. The home router rewrites the header of the packet, changing the private address of the device with its own public address. The packet is then sent to the public Internet by the router, which also keeps track of the combination of original private address of the device and destination address. When the router receives the response from the outside network, it changes back the address on the packet with the address of the device that initially request it. Doing NAT is a desirable practice in many local network, as it reduces the number of IP addresses used on the Internet.

Figure 5. Details on the Spoofer application (adapted from Beverly & Bauer, 2007)



Recruiting volunteers to run the test

The Spoofer Project relies on volunteers to run the test from within the network under test. This represents a big limitation not only for the availability of data, but also for their significance. In fact, having volunteers in the network might imply a certain level of awareness about the problem, as well as a certain level of technical qualification. Thus, it might be the case the tests are run more frequently on networks with operators already inclined to do proper security, leaving the worst networks in the shadow.

Researchers of CAIDA have promoted their application during conferences and seminars, in order to mobilize volunteers to run the Spoofer test. An alternative approach to increase the coverage of measurements of compliance is to use crowdsourcing marketplaces to recruit participants to download and run the test application for small monetary rewards. Crowdsourcing marketplaces are online platforms in which users are offered small compensation of few Euro to perform various types of micro-tasks (e.g. labelling images, translate short text or filling a questionnaire). Crowdsourcing appears a promising approach to gather volunteers to run the Spoofer test, as it might enlarge the geographical coverage of the measurements including networks that are otherwise difficult to test. The use of such platforms for network measurements has already been introduced in the literature and has been implemented for different cases, including the Spoofer test. Huz, Bauer & Beverly (2015) have successfully used Amazon Mechanical Turk to crowdsource measurements of deployment of IPv6, but failed to get participants download the Spoofer test (because Amazon's Terms of Service prevented users from downloading executable software for security reasons). As the Terms of Service of Mechanical Turk changed over time, a more recent study focused on the case of Spoofer, surveying different platforms (including Mechanical Turk) to gather participants to run the test (Lone, Javed, Korczynski et al., in press). The results of this second study were positive: with a budged of € 2000, 342 new ASes were tested (15% increase over the previous 12 months of tests).

2.2.5 Incentives analysis

After discussing the technical details of IP spoofing and its mitigation, we now turn to investigate the incentives to implement ingress filtering. In this section we focus on network operators, the organisations that run and maintain Internet's core technical infrastructure. As it might have emerged from the previous sections, network operators are the central actors in the mitigation of the problems related to IP spoofing, as it is their responsibility to implement source address validation.

The term *network operator* has a very broad meaning, and it is used to refer to a variety of organisation. First, there are *access providers*, companies that connect subscribers to the Internet. Often, access providers are not directly connected to the Internet backbone, but they receive access to it by *transit providers*. Next, there are hosting providers and data centres, that provide servers for web-hosting or online storage to both users and organisation (i.e. *content providers*), in order to further provide content or services on the Internet. Additionally, large enterprise and institution like universities might run their own infrastructure, representing another typology of network. Ideally, all these operators must have source address validation implemented (Internet Society, 2015).

Each category of operators can include very different type of organisations, not only in terms of priority given to security, but also in regard to their target customer, their size and their business model. Therefore, each individual operator will respond to a different subset of incentives, shaped by its specific market position (van Eeten et al., 2010). However, as the distinction between type of category is blurred (some large operators provide multiple type of services), so most of the incentives will be common among categories, though their effectiveness might vary.

In the following subsection, we discuss which factors determine network operators' incentives. After that, we will discuss the incentives for network operators to implement source address validation, and possible policy intervention to increase them.

Network operators' incentives

Given their central and essential position in the Internet ecosystem, it is not surprising that network operators are involved in the mitigation of many types of abuse, from blocking offensive contact, phishing, spam and malware in general, to ongoing debates on privacy protection and net neutrality. Hence, several studies have investigated network operators' in incentives.
We take as a reference point the framework designed by van Eeten et al. (2010), which investigate the role of ISPs in the mitigation of botnet, and we adapt it for the case of IP spoofing. The resulting model is shown in Figure 6.

The structure of incentives of a network operator is shaped by a variety of factors, which can be divided in two groups: *institutional* and *organisational*. Despite interrelated in a number of ways (e.g. factors in one domain may affect or enable factors in the other domain), the distinction between institutional and organisational factors is useful because organisational factors are typically shaped by the management, whereas institutional factors can be designed by policy makers (van Eeten et al, 2010).

A distinction is also made between *positive* and *negative* incentives. Other things being equal, a positive incentive encourage deployment of source address validation. For instance, cybersecurity laws and regulation create a positive incentive for operators to comply with them. Contrariwise, a negative incentive drives down security investment. An example of negative incentive is the costs associated with the anti-spoofing filters: costs of implementation, maintenance and training of personnel can discourage its adoption. Finally, some incentives can work both ways. A good example can be the costs of infrastructure expansion: not implementing anti-spoofing filters might result in bandwidth being consumed in DDoS traffic, which in turn would force investments in infrastructure expenditure.

Nonetheless, to implement ingress filtering it might be necessary to invest in new equipment, which may represent a disincentive to deploy (van Eeten et al, 2010).

Institutional factors:

- *Governance, regulatory, law enforcement*
  Undoubtedly, the regulatory context in which network operators are immersed plays a fundamental role in shaping their incentives. Proper Internet governance can design policy intervention to align actors' incentives, rewarding desirable behaviours and pursuing criminals. Likewise, the diligence and the reactiveness of law enforcement represent another important factor that might affect operators' incentives to adopt security measures.

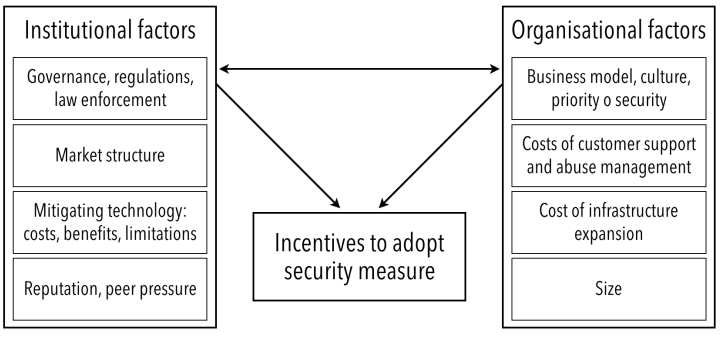
- *Market structure*
  The type and the structure of the market can contribute to operators' incentives too. For example, operators in highly consolidated markets with a handful of competitors might have lower incentives to adopt security measures when compared to their counterparts in more competitive markets.

- *Mitigating technology (costs, benefits, technical limitation)*
  Clearly, the characteristics of the mitigating technology is an important factor. Specifically, its costs and benefits (often articulated in terms of Return on Security Investment) and its technical limitations contribute to determine its.

- *Reputation*
  Incentive are not only monetary. Non-financial factors such as reputation and brand damage can be powerful motives to prompt the adoption of a particular technology. In some cases, a drop in reputation can have serious financial implication, for example in terms of stock price.

- *Peer pressure*
  Similarly to reputation, peer pressure can represent a strong incentive for network operators. A good example is the case of blacklists, lists of elements (IP addresses, URLs, domain names...) blocked from accessing some resources as they have been reported due to an abuse. Blacklisting an IP address can produce significant economic impact for its operator, as it might increase the cost of customer support.

Figure 6. Main factors contributing to operators' incentives (adapted from van Eeten et al., 2010).



Organisational factors:

- *Organisation norms and values (business model, culture, priority to security)*
  As security comes to a cost, it seems rational for network operators to tolerate a minimum level of cyber risk. The business model and the organisational culture of a network operator contribute to determine the priority given to security and the degree of security attention. On one hand, there are some highly vigilant operators who make their attentiveness a key selling point. On the other hand, there are the so-called *rogue operators*, who deliberately allow questionable practices for financial gain (e.g. spreading spam and malware or other fraudulent schemes) (van Eeten & Bauer, 2008). The rest of operators swing between these extremes, with a degree of security vigilance that is the result of a mixed set of incentives.

- *Size*
  The size of a network operator (e.g. the number of subscriber or the number of IP prefix managed) is an important variable that can mediate the effectiveness of other factors. For example, large operators might be more concerned with maintaining their brand image. Similarly, small operators might have less expertise and may dispose of a smaller capital to invest in security.

- *Cost of customer support & abuse management*
  The cost of maintaining an efficient abuse department to address complain, together with the cost customer support represent another factor that can affect the incentives of an operators to adopt a particular security measure. For example, not implementing a security measure might results in a high number of compromised customers who seek support. In this case, the cost of customer support represents a positive incentive, as it induces operators do invest in security. However, providing extensive support to customers also drives up costs, as it might require communication with a large number of users. In this second case, the cost of customer support represents a negative incentive.

- *Infrastructure expansion*
  The cost of infrastructure expansion can have both a positive and a negative impact on operator's incentives to invest in security. Networking equipment is generally expensive, and operators can be not inclined to invest in security measures that require additional equipment. However, repeated abuses due to lack of security might consume available resource, thus requiring investments to expand the current infrastructure.

After having provided an overview on the factors that contribute to network operators' incentives, we turn now to discuss in more details their incentives to do source address validation.

Incentives to deploy source address validation

In the next two sections, we explore the factors that play a major role in the decision to deploy source address validation. In particular, we discuss the results of a round of preliminary interviews conducted to investigate network operators' incentives. The interviews were designed on the base of model presented above, and took into account a number of input from available literature on the topic (mainly

Beverly et al., (2009) Damas (2008), McConachie (2014) and Internet Society (2015)). Details on the methodology and the analysis of the interview are reported in Appendix 2.

As a starting point, it is important to highlight the presence of externalities in the problems related to IP spoofing, especially when considering reflection and amplification DDoS attack. Despite plausible, it might not be precise to claim network operators create negative externalities by not implementing SAV. It is certainly true that the costs of a DDoS attack originating in a network which does not prevent spoofing are paid by the victim of the attack (clearly a negative externality), but the non-deployer network is not the actual source of the externality (i.e. is not the "polluter"). Instead, it is possible to claim that by implementing SAV network operators generate positive externalities. As a matter of fact, the main beneficiary of the deployment of SAV are other networks and the Internet ecosystem at large. This is the main obstacle to the diffusion of anti-spoofing filters: the misalignment of incentives between who is responsible for security and who enjoys its benefits. Network operators may adopt all best practice but still receive anonymous, malicious incoming traffic from third parties who do not filter properly (Beverly et al., 2009).
Furthermore, from the point of view of the community, the utility of doing SAV depends on the number of network implementing it. In fact, the benefits of deployment are limited until it is widely adopted (i.e. until the majority of the networks can be used to launch DDoS attack) (Baker & Savola, 2004). The lack of substantial benefits and this type of network effect may represent a disincentive for operators to implement SAV.

This situation might resemble the tragedy of the commons: selfish network operators want to save money and effort by not invest in anti-spoofing and, as a result, the overall level of security of the Internet decreases (McConachie, 2014). Moreover, it seems that operators, even the most vigilant, do not perceive IP spoofing as a priority. Though it is a critical problem from a societal point of view, it is not a specific focus problem for operators. *"It is less than spam, and it is not directly harmful for us"*, confessed one participant during the interviews (who, however, continues *"but you do it* [ingress filtering] *because it's the right thing to do!"*). As Hardin's herdsman does not consider an extra cow on the common field much of a problem, so some operators might think to spoofing as "not a big deal".
Nevertheless, the argument of the tragedy starts to crumble when the benefits of doing SAV and the actual costs of not doing it are further investigated.

When the benefits of doing SAV are reframed from *preventing other networks from being attacked* to *"preventing your customer from being used as a launch-pad for attacks"*, it appears immediately clear that there are deeper implications. Albeit there is not any liability on the network in which a DDoS originates, its reputation can suffer. For instance, consider the inconvenient position of an operator having to explain to the victim of a DDoS that the attack was not actually his responsibility, despite he could have prevent it (Damas, 2008). No operators want to be involved in such attacks, if only because attacks attract the attention of more attackers. Moreover, bandwidth does not come for free: operators pay for the spoofed packets egressing their network during a DDoS (McConachie, 2014).
Beside preventing DDoS attacks, implementing SAV brings additional side benefits to network operators. Firstly, deploying SAV filtering preclude attackers the possibility of minor, yet troublesome exploits (e.g. spoofed management access to networking equipment) (Baker & Savola, 2004). Secondly,

having only traffic coming from an authenticated source is highly beneficial both in case of ordinary traffic analysis and in case of forensic investigations after an attack. In addition, dropping packets with a non-valid source address at the edge of the network saves bandwidth and resources, contributing to maintain the core of the network clean.

It is tempting to see the diffusion of SAV in terms of tragedy of the commons, but the analogy does not hold up. Implementing anti-spoofing filters is in everyone's interest, even of selfish operators (McConachie, 2014). Nonetheless, it appears that the incentives to deploy SAV are low, partially due to the externalities, and partially because operators might feel that these benefits do not overweight the costs of deployment. But what costs are associated with the implementation of SAV?

In general, the costs of deployment are minimal *("and perhaps should not be perceived as extra costs, but as part of the daily part of ISPs' daily job as a maintainer of the network…to have good hygiene on your BGP configuration"*, noted a participant in the interviews). Surely, the efforts to deploy and maintain ingress filters depend on the particular topology of the network implementing it, and grow as the complexity of the network increases. However, at least for the case of most access providers, these costs are marginal.
Today, most part of routing equipment include ingress filtering by default. There might be instances of old equipment that does not support uRPF, and so the deployment of ingress filtering would necessitate replacing the equipment, but eventually, this older equipment will gradually have to be replaced. When uRPF cannot be implemented, manually deploy filters can be complicated. A valid concern, in fact, refers to the manual maintenance of large access control lists prescribed by BCP38 (Beverly et al., 2009). In cases of fast growing networks or very large networks, *"this is definitely a hassle"*, confessed an operator, who also added that the diffusion of cloud computing and network virtualisation further complicate maintenance.
On top of the costs just discussed, another consideration is worth mentioning, as emerged during the interviews. Implementing ingress filtering, as implementing any new feature on a network, is a risk. It might not be expensive and technically infeasible, but it is definitely time consuming and error prone. Setting up and testing the configuration from scratch might take up to months, and necessitate good knowledge about the internal structure of the network. Moreover, some manual implementations require attention, especially because the risk of cutting out customers is high, and is not worthy for a functionality with a limited profitability. Lastly, in some cases, uRPF can lead to dramatically drops in performances, which may force operators to disable it.

The situation is further complicated by the incomplete information available on the network that allow IP spoofing, and by the difficulties in measure compliance. On one hand, the absence of a clear picture on which networks deploy anti-spoofing and which do not, limits the ability of governance bodies to design tailored interventions (Beverly et al., 2009). On the other hand, the impracticality of revealing the lack of filtering contributes to operators' incentives not to deploy, as they can pass unnoticed. Considering its global scale, addressing the challenges that spoofing poses without reliable and complete data is lake shooting in the dark (Internet Society, 2015).

Finally, there are zones of both the virtual and physical word with a high level of DDoS activity and cybercrime in general. Operators in these areas might not be moved by arguments about network hygiene or reputation damages. What is worst, their careless behaviour does not only represent additional challenges to address, but affects also the decision making of other operators. Social loathing and pessimism exists also in Internet security, and some operators may wonder about the worthiness of their contribution when there are other networks performing much worst in terms of security. However, the presence of some "bad apples" does not represent a valid justification to lower security standards.

In conclusion, SAV adoption appears as a challenge due to the misaligned incentives between who spend in security and who receive the benefits. Operators' incentives to do SAV are determined by a wide range of factors, some clearly having a positive influence, some with a negative influence, and some that might work both ways. Moreover, some factors can have a combined effect, in the sense that one factor can amplify or reduce the perception of other factors.
In particular, the benefits of doing SAV and the costs of deployments have been discussed, and are schematised in Table 1. It has been showed that implementing new functionality on a network always carries some level of risk. We also highlight the presence of non-technological barriers, like externalities and incomplete information, and we pointed out that situation of FUD may further contribute to lower the incentives to do SAV. Therefore, a spontaneous question arises: what can be done to address these challenges and increase the incentives to implement SAV?

Table 1. Network operators' incentives to deploy filters.

Positive incentives	Negative incentives
Prevent propagation of DDoS attack	Lower performances, risk of configuration errors
Good reputation	Cost of infrastructure expenditure
Improve network hygiene	Not perceived as a problem/no regulation
	Lack of information/FUD

Possible interventions

As the implementation of SAV appears to be primarily beneficial for networks other than the one deployer, so the incentives of operators are misaligned. The presence of externalities and information asymmetries would suggest that an intervention might be necessary to further incentivise network operators.

A first, instinctive approach can be regulatory intervention. National telecommunication authorities can require operators to implement anti-spoofing measures. During the interviews, this type of regulatory approach has emerged as a potentially powerful incentive for operators. However, the only case we are aware of is represented by Finland, where operators must:

"prevent in their IP interconnection interfaces such IP traffic to the operator's network where the source address of a received IP packets 1) belongs to an IP address space that the telecommunications company itself administers or advertises, 2) belongs to an IP address space that is reserved for non-

public use, or 3) do not belong to routes advertised by a telecommunications company that conveys traffic to other telecommunications companies." (FICORA, 2015:3)

Despite regulatory intervention might move a strong incentive for operators to comply, it also hampered by a number of limitations. In fact, the process of establishing such regulation might be long and its enforcement can be costly (Anderson & Moore 2007). Moreover, the successful implementation of this type of regulation is bounded by the availability of data and on their reliability.

A second type of approach to regulatory intervention is represented by self-regulation. In fact, during the interviews, it has been proposed that "responsible" network operators, who already deploy anti-spoofing, can require their peers (i.e. other networks with which they exchange traffic, freely and for mutual benefit) to have ingress filters in place, or else "de-peering" them. However, given the mutual benefits of peering relationship, operators can prefer having peers rather than preventing spoofing, especially if operators do not perceive IP spoofing per se as a threat.
Another category of actors that can have more leverage in establishing a similar self-regulation are transit providers, organisations that provide connectivity to other operators. Contrary to peering agreements, transit agreements are priced: operators pay transit providers to receive a certain amount of bandwidth. Thus, it seems reasonable that transit providers can require their customers to deploy anti-spoofing, or charging a premium to non-deployer networks. After all, in case of a DDoS originating from one of their customers, transit providers may carry part of the spoofed packet used to initiate the attack. Yet, as transit providers are paid depending on the amount of bandwidth consumed by their customers, they might have an incentive to carrying as much traffic as possible, even if malicious. Clearly, this might produce another, undesirable, incentive problem.

Instead, a promising initiative of self-regulation is represented by the Mutually Agreed Norms for Routing Security (MANRS). MANRS is a community-driven initiative, led by the Internet Society, aimed at repairing the cracks in Internet's routing infrastructure. Participants in the MANRS community agree to make a social contract, and to give their contribution to the good of the commons. This contribution is articulated in three fundamental actions (plus one advanced):

1. *prevent propagation of incorrect routing information;*
2. *prevent traffic with spoofed source IP address*
3. *facilitate global operational communication and coordination between network operators*
4. *facilitate validation of routing information on a global scale.* (MANRS, 2015: 3)

By bundling the prevention of spoofing together with other actions, the MANRS initiative aims at integrating the limited benefits of anti-spoofing with those of other security measures. However, as the initiative was born is 2014, there are less than two hundred networks participating.
It is definitely a long run for MANRS to reach a large number of participants. Yet, the idea of creating a community of responsible operators appears promising. Eventually, *MANRS compliance* may become a sort of security certification that can be requested as a sign of trustworthy to operators and enterprises, who can in turn diffuse it in their supply chain.

Another attracting approach to increase network operators' incentives to implement anti-spoofing involves the use of reputation and peer pressure. Disclosure policies aimed at publicly revealing the lack of anti-spoofing might be beneficial for the reason briefly presented in Section 2.1.4 (namely: being a sanction for non deployer operators, serving as a deterrent for other operators, and establishing that allowing IP spoofing is not tolerable). On one hand, reputation might work because of the incentive of the majority of operators to maintain a good brand image. On the other hand, it is reasonable to assume that deployer operators, who have already given their contribution to Internet security by implementing ingress filtering, might have an incentive to require other operators to do so as well. Moreover, given the information asymmetries, policies aimed at disclosing information can be particularly effective.
Over the last years, CAIDA's Spoofer Project has collected measurements of deployment of ingress filtering worldwide. The results of the measurements are published on CAIDA's website with different level of aggregation. When the test identifies a non-deployer network, researchers of CAIDA report the lack of ingress filter to the network administrator, in order to instigate remediation. The effect of their notification is encouraging: 15-20% of the contacted operators remediated (Claffy, 2017). Nevertheless, it seems that the majority of non-deployer networks ignored the notification they received.
Hence, a series of questions emerges: how to further increase the remediation rate achieved by researchers of CAIDA? Does reputation represent an effective incentive for operators to implement ingress filtering? In what ways can we engage the community of operators to create pressure on the networks allowing spoofing? In the coming chapters, these questions will be investigated. Specifically, the next chapter will discuss abuse reporting and vulnerability notifications, with the aim of building a theoretical model on which to base our hypothesis. Our goal is to design an experiment to strategically disclose information about non-deployer operators, in order to evaluate the possibility to use reputation and peer pressure as a means of compliance.

2.3 Conclusions

This chapter presented an insight on cybersecurity problems from an economical perspective. In the first part of the chapter, we introduced the field of the economics of cybersecurity, pointing out that many security problems can be reformulated in light of microeconomics concepts. In fact, despite an initial technical appearance, the lack of security may be symptom of perverse incentives. In the fight against cybercrime, technological solutions cannot represent the only weapon for security defenders, whose arsenal should also include policy interventions to address actors' incentives. However, no *silver bullet* exists for cybersecurity: every solution has its limitations, and might produce different effects if applied to different problems.
Next, we shifted our attention to the case of IP spoofing. We have shown that economic barriers like externalities and information asymmetry contribute to further complicate the mitigation of a problem already technologically complex. Operators are presented with a demanding solution to a problem they do not perceive as an imminent threat. So, what can be done to increase their incentives to adopt of anti-spoofing measures? After having reviewed some possible solution, we argued that interventions aimed at alleviating the information asymmetries can be beneficial. In particular, we have considered reputation and peer pressure as potentially effective incentives to increase adoption of anti-spoofing measures. In the next chapter, dealing with security notification, we will discuss in more details the role of reputation and peer pressure in increasing the effectiveness of vulnerability notifications.

Chapter 3:
A framework for public notifications

The aim of this chapter is to design a conceptual framework that will be used to represent the variables under investigation and their relations. On the base of this model, we will state the empirical hypothesis that will be tested during our experiment. In order to identify the main variables, to explain their relations and, thus, to create such model, we shall start by reviewing the available literature. In particular, we begin discussing the importance of sharing security information (Section 3.1). Next, we focus on vulnerability notifications, to review the impact of several empirical studies (Section 3.2). In Section 3.3, we present insights from other literatures on the effects of disclosing social information on people and corporates' behaviour. Finally, the conceptual model is designed in Section 3.4

3.1 Sharing security information

In the previous chapter, we argued that the cooperation among defenders represents a necessary condition to improve Internet security. Such cooperation typically takes the form of sharing security information (Gordon, Loeb & Lucyshyn., 2003; Gal-Or & Ghose, 2003; Schecter & Smith, 2003). In fact, collecting and analysing information about successful and unsuccessful cyberattacks can improve our understanding of attackers' modus operandi and choice of targets. It could contribute to improve our capability to prevent breaches and detect anomalies, as well as to design better strategies to response to incidents and recover from them. In particular, when this information is shared among defenders, it can prevent organisations from falling victim of cyberattacks already experienced or blocked by other parties.

For this reason, the turn of the millennium saw the rise of Information Sharing and Analysis Centers (ISACs), non-profit organisations aimed at facilitating such knowledge sharing process. ISACs collect and disseminate information about system vulnerabilities, threats and attacks among its members. Ultimately, by encouraging continuous improvements in security, ISACs also contribute to increase the demand for security products (Gal-Or & Ghose, 2003).

Sharing and disclosing information about security incidents can have both a negative and a positive effect on the disclosing organisation.

As a results of the disclosure of security breached the market value of an organisation may drop (Campbell et al., 2003; Acquisti et al 2006). It has also been reported that IT executives showed more concern about the effect of online frauds on customers' trust and confidence in e-business than about the actual financial costs (Gal-Or & Ghose, 2003).

The flip side of the coin is that information sharing is beneficial to both the individual organisations involved and to the whole Internet ecosystem. Individual benefits can be direct or indirect. For instance, identifying and remediating vulnerabilities represents a direct benefit as it prevents future breaches. Instead, a better security reputation is an example of indirect benefit, as it can indirectly lead to

increases in sales. In fact, sharing security information with law enforcement can be seen as a strong message to consumers that an organisation devotes care to security and that it takes actions to contribute to the good of the Internet ecosystem (Schecter & Smith, 2003).

Information sharing has been under investigation for a while in the economical literature, though not with the specific focus of information security (Gal-Or, 1985; Shapiro, 1986). Since the early 2000, scholars have turned the attention to the case of information security, in order to contribute to the emerging literature of the economics of security. Gordon & Loeb (2002) develop an economic framework to determines the optimal amount of security investments to defend a given set of information. On the base of that model, Gordon, Loeb and Lucyshyn (2003) further examine the case of information sharing, concluding that sharing is beneficial both to individual organisations and to the total welfare of the ecosystem. Nevertheless, they also argue that without an appropriate incentive mechanism, organisations may free ride on other's expenditures.

While Gordon and colleagues focus on the overall level of information security, Gal-Or and Ghose consider the impact on the individual organisation. Gal-Or and Ghose design a model to study the effect of market characteristics (e.g. competitively and size of organisations) on the level of information shared. In addition, they formalise also how information sharing influences the demand for the final product, the production costs as well as the cost of security investments and the cost of information sharing (Gal-Or & Ghose, 2003).

3.2 Vulnerability notifications

One particular form of information sharing is vulnerability and abuse reporting, where one party notifies another one about potential abuse, requiring taking action against it. In this section, we propose an overview of the most relevant literature dealing with the notification of security vulnerabilities. We review empirical findings that aimed at understanding the effect of reporting vulnerabilities and abuses on the remediation of compromised resources.

Recently, the security community has turned its attention to the effectiveness of vulnerability notifications. Jhavieri et al. (2017) propose a framework model of the abuse reporting infrastructure, identifying three intervention strategies: direct remediation, intermediary remediation and third party protection. Direct remediation happens when the notification is addressed directly to the owner of the compromised resource. In intermediary remediation, instead, an intermediary is contacted with the goal of instigate remediation by the resource owner. Third party protection occurs when the notification is sent to a security vendor, who leverages it in order to protect a third party, (usually a client). For each strategy, Jhavieri and colleagues describe the actors involved as well as the incentives behind their action. In particular, they investigate the incentives for researchers and volunteers to collect abuse data, for intermediaries to act on abuse notification and for affected resource owners to remediate.

Cetin, Ganán, Korczynski & van Eeten (2017) conduct large scale notification campaigns to test the effect of different variable on the remediation of a vulnerable configuration of DNS server, exploited during zone poisoning attacks. First, they analysed three different channels to reach affected party

(directly reaching nameserver operators, and intermediary remediation contacting domain owners and network operators), concluding that none of them provided reliable contacting information. However, for those entities that were reached, notification positively influenced remediation. Secondly, the authors studied the incentives mobilized with each channel. They found that direct remediation (by nameserver operators) leads to better results than contacting an intermediary (the domain owners), who, on paper, should have better incentive to remediate. Finally, they built a demonstration website to inform about the vulnerability, but found little engagement of participants with the website.

Previous work by Stock, Pellegrino, Rossow, Johns, & Backes (2016) also surveyed existing communication channels, getting to a similar conclusion of Cetin and colleagues: notifications have a positive impact on remediation, but finding reliable contact information is a major obstacle to this process.
To tackle this challenge, Stock, Pellegrino, Li, Backes & Rossow (2018) further analysed what technical and human aspects are roadblocks to effective vulnerability notifications. They discuss the shortcomings of email-based notifications, such as problems related to anti-spam filters, lack of trust by recipients, and hesitations to remediate despite awareness. They also conduct a large-scale notification experiment to reach the owners of more than 24,000 domains, probing possible alternative methods beyond emails, including social media and phone, but failing to identify an effective channel.

Zhang, Duan, Liu & Yao (2017) extended previous research to the scope of an ISP. They focused on a Chinese ISP to compare the effectiveness of three notification methods to report vulnerability to customers. They show that notifications improve vulnerability remediation, though their effectiveness depends factors like the channel reliability, the characteristics of the vulnerability, contact's authority and technical capability. They also noticed that the number of notification times is not important for remediation if the notification method is not appropriate.

The positive impact of notification has also been proved by Li, Ho et al. (2016) who detailed describe the effect of abuse notification more than 700,000 hijacking incidents detected by Google Safe Browsing and Search Quality. The likelihood of remediation increased of 50% when webmasters are directly notified. At the same time, they observed that notifications shorten the length of infection by 62%.

Cetin, Jhaveri, Gañán, van Eeten & Moore (2016) focused the role of sender reputation in the notification process and, ultimately, on the cleanup rate of Asprox botnet. Despite a significant effect of notification over a control group, they find no evidence that the sender reputation affect cleanup. Furthermore, they show that the minority of participants engaged with a demonstration website were more likely to remediate.

Li, Durumeric et al. (2016) investigated further factors that might affect the remediation rate for different security vulnerabilities. Their most interesting findings are that: contacting directly the resource owner via WHOIS record, a standard protocol to retrieve information on Internet resources, lead to better remediation than CERTs; as well as providing very detailed information on the vulnerability (over terse messages) improves the chances of remediation. Moreover, they found no significant evidence of remediation in the case of notification about DDoS amplifier misconfigurations.

Prior work of Kührer, Hupperich, Rossow & Holz (2014b) also investigated the mitigation of amplified DDoS attacks. Among other protocol that are exploited for amplification, they focus on the misconfiguration of NTP and collaborated with the security community in a large-scale campaign to reduce the number of misconfigured servers by 92%.

We have already mentioned the work of researchers of CAIDA to instigate compliance with anti-spoofing best practices, in order to reduce the Internet's susceptibility to amplification DDoS attacks. They have promoted the Spoofer Project for more than a decade, developing a measurement application to test compliance with SAV. When the application reveals the lack of SAV, they sent manual notifications to the operator of the network. They report a moderate remediation rate around 18% (one operator on six remediates, one on five for operators in English speaking countries) (Claffy, 2017).

Researchers have also investigated whether disclosing vulnerability information publicly may produce better outcomes in terms of remediation (Arora et al., 2004). Arora et al. (2010) found evidences that the public disclosure can accelerate patch release.

Moore and Clayton (2011) analysed the rate to which phishing website are re-compromised after an initial clean-up. They conclude that around 17% of phishing website are re-compromised in one year. However, they also notice that websites reported on a public blacklist are re-compromised less often than those only reported within a closed community, concluding that publicly revealing of vulnerability information can actually aid defenders.

In the mitigation of botnets, Gañán, Cetin & van Eeten (2015) showed that different notification regimes influenced the lifetime of ZeuS C&C servers. Researchers concluded that publicly accessible C&C servers were mitigated 2.8 time faster than non-publicized. Moreover, they pinpoint to location and type of hosting as two important factors in the C&C infrastructure lifetime.

Finally, a group of researchers has focused on the use of public disclosure to instigate remediation out of reputation concerns (Linden et al., 2012; Tang et al., 2013; He, Lee, Han & Whinston, 2016).
In particular, Tang et at. (2013) described the mechanisms behind information disclosure in terms of social comparison theory, explaining elements such as status, shame and fame. By publicly disclosing information about organisation outbound spam, they achieved a reduction of outgoing spam of 15.9%, noticing however that networks originating the greatest amount of spam were indifferent to the rankings, suggesting that they were unconcerned with their public reputation.
In a prosecution of that study, He et al. (2016) observed a reduction in spam also in large spammers, and found evidence of significant peer pressure among organisations.
These papers are framed in the larger research project of Spamrankings.net, which over the years has been presented in many conferences resulting in different papers. The general objectives of the Spamranking project is to study if public ranking of spam can be an effective mechanism for encouraging firms to reduce outbound spam. This group of authors introduced Spamranking at the RIPE conference of 2010 (Quarterman, 2010), proposing their approach based on systematic public rankings to improving Internet security. Over the years, they show how major botnet takedowns (*Ozdok* or *Mega-D, Rustock*

and *Grum*) did not have much effect on spamming or security beyond temporary short-term effects, arguing for the use of reputational rankings as a new and more encompassing approach to Internet security (Quarterman & Whinston 2010; Quarterman, Sayin, & Whinston, A 2011; Linden et al. 2012).

3.3 Social information and public disclosure

In this section, we focus on the use of social information and reputation to steer individuals and organisations towards desirable behaviours. In particular, Section 3.3.1 introduces social comparison theory, a psychological framework originally developed by psychologist Leon Festinger (1954). Festinger suggests that, when information on others is provided, individuals evaluate themselves in comparison with their peers and, in turn, this comparison affect people's behaviour. Next, in Section 3.3.2, we focus on how reputation can be used to instigate organisations to comply with norms and regulations from a policy-making point of view. These insights are important for our research because we aim at increasing the effectiveness of vulnerability notifications by disclosing information about compliance with anti-spoofing measures. Therefore, the aim of this section is to discuss the mechanisms by which information disclosure can increase incentives to act on the notification.

3.3.1 Social comparison and pro-social behaviour

In a social community, individuals show the tendency to evaluate their abilities, opinions and performances by comparing themselves to others (Festinger, 1954). When information on other people's behaviour is available, such social comparison is accountable for shifts in individual's opinions and for an increase in motivation and competition in the community, especially in ambiguous and confusing situations (Festinger, 1954; Allen & Wilder, 1977; Suls, Martin & Wheeler, 2002).
Over the years, psychologists have further extended social comparison theory by distinguish between cases in which the comparison is made with an individual considered worse off (i.e. downward comparison), or with an individual considered better off (i.e. upward comparison). Individuals engage in downwards comparisons as a defensive mechanism to feel better about themselves and increase self-regard (Wills, 1981). In contrast, upward comparison may result in lower self-regard, envy and depression (Tesser, Millar & Moore, 1988; Wheeler & Miyake, 1992). Nonetheless, upward comparison may also open the door to self-improvement and self-enhancement, as people are moved by the desire to be part of the elite (Collins, 2000). In this case, *"upward comparison with superior models can provide hope and inspiration"* (Suls, Martin & Wheeler, 2002: 161).
Hence, social comparison might cause an increase in competitivity among peers, in light of people's concern for social status and reputation (Forsyth, 2000). People's efforts to build and maintain a good reputation can originate both from intrinsic reasons, as status is an inherent human characteristic, and instrumental reasons, as it can enable better future opportunities (Postlewaite 1998).

The availability of information about individuals' behaviour in the community represents a prerequisite for social comparison. Recently, social comparison has been used as a means to instigate *pro-social behaviour*, in what has been defined *conditional cooperation*: individuals are more willing to contribute when they know that others are contributing (Frey & Meier 2004). For example, Satio (2011) proposes that people feel ashamed about decisions that do not maximise the payoffs of others, but only if the

decision is made in public. Likewise, Dillenberger and Sadowski (2010) suggest that the outcome of a person's decision-making changes when it is observed by someone directly affected by that decision, and they thus include shame as a moral cost to a person's utility.

Hence, social comparison and social information can be useful tools to address social loafing (i.e. an individual tends to contribute less towards a goal when he works in a group than when he works individually). Arguably, social loafing occurs due to reduced individual incentives and lack of coordination (Karau & Williams 1993). Both these are present in many cyber security problems: low incentives arise from externalities and coordination losses are the consequence of the cost of security efforts (Tang et al., 2013). Imposing reputation damage by disclosing relevant social information may work as a binding force to fight social loafing (Akerlof, 1980). The social norms created with the disclosure of social information contribute to pro-social behaviour in two ways: by having a focusing influence (i.e. a norm impacts behaviour only if individuals' attention is drawn to it), and by having an informational influence (i.e. the more people are seen behaving consistently with a norm, the stronger the impact the norm has on other individuals) (Krupka & Weber 2009).

Increasingly, social comparison and social information are used in experiments to trigger pro-social behaviours. We have already discussed how Tang et al. (2013) used social information to reduce organisations' outbound spam in the previous section. Other empirical findings have shown the positive effects of providing social information in the case of charitable donations (List & Lucking-Reiley, 2002; Frey & Meier, 2004; Shang, & Croson, 2009), political participation (Margetts, Escher & Reissfelder, 2011), household electricity and water savings (Ek & Söderholm, 2010; Grønhøj & Thøgersen, 2011; Ferraro, Miranda, & Price, 2011; Ferraro & Price, 2013), contribution to online communities (Butler, 2001; Beenen, Ling, Wang et al., 2004; Ludford, Cosley, Frankowski et al., 2004; Chen, Harper, Konstan et al., 2010).

These studies also present a partial overlap with research aimed at influencing and improving individuals' decision making. Behavioural economists have recently focused on *nudging*, which consists in designing a system in a way that users are (unconsciously) prompted to prefer one choice over another. An explicit example of nudging is the card presented in the many hotels' toilettes, which nudges guests to pay attention before having towels or sheets needlessly washed.
Nudging has been around since the 90s as a mechanism to influence groups and individuals' decision making. Nudging techniques take up from the concept of *weak, libertarian*, or *soft paternalism* (Thaler & Sunstein, 1999), and have been widely deployed, for example, to prompt users to take better decisions about their privacy (Acquisti, 2009).

3.3.2 Public disclosure as a mean of compliance

Reputation and corporate behaviour

In very simple terms, reputation can be defined as the confidence that a person will keep his promise in the future, on the base of his behavior in the past (Macaulay, 1963; Ellickson, 1994; van Erp, 2011). In this simple formulation, the concept of reputation applies both to individuals and to organisations. Fombrun & van Riel defined corporate reputation as the *"collective representation of a firm's past*

actions and results that describes the firm's ability to deliver valued outcomes to multiple stakeholders" (Fombrun and van Riel, 1997: 10).
Corporate reputation and trustworthiness represent a type of capital that enables an organisation to grow and to build a robust market position. A positive reputation is the key to open the doors to new consumer markets, new business opportunity and investors, and hence to achieve financial objectives. However, reputation does not only account for financial goal, but is also valued intrinsically, for example in terms of self-esteem and confidence (Fombrun and van Riel, 1997).

According to Gunningham, Kagan & Thornton (2004), reputation constraints organisations to operate in according to a *social licence*, in the sense that they are required to act in order to meet the expectation of society and, in particular, of stakeholders like customers, employees, investors, peers, the media and non-governmental organizations (NGOs). In response to these expectations, corporate behaviour is driven by a fundamental concern for good reputation. Decades of empirical research showed consistent evidences of the positive effect of reputational concern on corporate behaviour (Macaulay, 1963; Ellickson, 1994; Chatterji & Toffel, 2010).

Given the positive effect of reputational concern on corporate behaviour, it is not surprising that regulatory interventions aimed at disclosing organisations' wrongdoings and expose them to reputational damage can induce desirable behaviours. Disclosure policies, or *policy-by-revelation*, offer an alternative type of governance to the traditional surveillance and coercion approach (Florini, 2008; Gupta, 2008). The disclosure of the names of organisations with rogue behaviours is also known as *"naming and shaming"*, a controversial term to which the more neutral word "disclosure" is often preferred (Financial Service Authority, 2008).
While financial sanctions might be perceived as a weak signal or a weak threat, and their deterrent effect can be limited (Thornton, Gunningham & Kagan, 2005), disclosure policies can have a better impact because, in addition to the financial costs, they may produce bad publicity and shame in the eyes of investors, peers and customers (van Erp, 2011).
Specifically, reputational sanctions can produce multiple effect: first, they may induce financial damage due to drops in sells or stock price; second, they are negative publicity; third, they can increase organisations' awareness of duties and obligation, and lastly, they increase transparency and empower consumes (van Erp, 2011). In the follows sections, we briefly describe the effects more relevant for our intervention:

Shame and bad publicity

The effectiveness of reputational sanction arises from the expectation that various stakeholders will react to the disclosure by avoiding firms whose wrongdoing is revealed. Disclosure can lead to a reduction of sales, loss of business opportunity and drop in market share and stock price. Moreover, organisations fear bad publicity and public shame. People, as well as organisations, strive for good reputation because they value being considered reliable, trustworthy and respectable. Hence, fear of losing status, trust and legitimacy represents a valid incentive for organisation to comply with regulations (van Erp, 2011).

Establishing pro-social behaviour

Reputation losses can be considered as direct incentives for organisations to comply with norms. However, public disclosure can also have indirect effects. In fact, apart from being a threat message, reputational sanctions carry a message about appropriate and desirable behaviours. Black, Hopper & Band argues that disclosure policies provide examples of good and bad practices, inducing *"reasoning by analogy"* that can prompt compliance in situation of uncertainty (Black, Hopper & Band, 2007: 201). Moreover, publicly exposing wrongdoings highlights the unacceptability of the behaviour (Braithwaite, 1989). Hence, disclosure policies can contribute to raise awareness about the issues that requires regulatory attention (Thornton et al., 2005). Finally, disclosure policies represent an implicit reminder that compliance is worthwhile and that rogue behaviours are penalised (Thornton et al., 2005).

Increase transparency

Disclosure policies are also aimed at increasing market transparency and alleviating information asymmetries between customers and organisations, by making information more accessible (Gunningham, Grabosky & Sinclair, 1999). Therefore, public disclosure empowers consumers, enabling them to play a more active role, and resulting in a democratisation of the power dynamics of the market (Gupta, 2008). Market pressure from peers and consumers represent a powerful incentive to comply with regulations, because it rewards compliance and at the same time penalises rogue behaviours.
Nevertheless, empowering consumers does not always generate this desirable engagement. Hutter and Jones (2007) suggest that consumers may undergo information overload, implying that people will not react as expected because they cannot process the information presented to them. In these cases, disclosure should be addressed to intermediaries in a better position to exercitate pressure on organisations, for example the media and NGOs. In this regard, van Erp (2011) reports the outcomes of the decision of the Dutch Health Inspectorate to disclose the quality of hospital care in the Netherlands. Though the goal of this disclosure policy was to empower patients to make more informed decisions, they did not react as expected. Nevertheless, the initiative successfully engaged patient organisations, who, instead, started pressuring on hospitals.

3.3.3 Final thoughts

All in all, in this section we have introduced social comparison theory, which explains how individual behaviour is driven by the availability of social information, and we have discussed the use of disclosure policies for instigating compliance. The reason of this digression is that we want to take advantages of the mechanisms just explained to improve the effectiveness of vulnerability notifications.

As discussed in Chapter 2, security decisions are driven by economic incentives. In particular, we showed that the case of IP spoofing is affected by economic problems that might result in perverse incentives (at least in the case of the 30% of operators found non-compliant). In this situation, simply reporting the problem to operators might not be enough to induce remediation. For example, the notifications sent by researchers of CAIDA to non-compliant operators achieved a remediation rate of roughly 20%, and though there might be additional factors that contribute to this low result (e.g. retrieving the right contact to notify, or just valid reasons for operators not to comply), we suppose that most of the operators simply ignored the notification. Therefore, we seek to craft our notifications to make the

message as compelling as possible. In this regard, disclosing information about compliance with anti-spoofing best practices seems a good approach to incentivise remediation.

The results presented in Section 3.3.1 suggest that releasing relevant social information is a valid way to tackle situations of individual under-contribution to common goods. Moreover, in Section 3.3.2 we extended this logic to the case of corporations, showing that public disclosure can create additional incentives for organisations to comply with norms and regulations. Thus, we aim at improving the effectiveness of vulnerability notifications by not simply reporting the problem to affected operators, but also by disseminating information about the number of networks found with and without anti-spoofing filters. In particular, we believe that highlighting that the networks lacking filters are a small fraction of the total number of network tested might nudge non-compliant operators to take into account our notification, and, ultimately, to deploy anti-spoofing measures.

The information we will release is already publicly available on CAIDA's website. Our intervention aims to aggregate this information, increase its visibility, and provide it to operators as a targeted feedback. To be more precise, our aim is to test two interventions. In the first, information about which operators lack anti-spoofing and which operators instead deploy it will be shared only with non-compliant operators as a private notification, to see if the availability of this type of information may induce remediation (e.g. by stimulating upward comparison). In the second, we will make our aggregated information publicly available, in an attempt to create additional incentive out of reputation concern and peer pressure. In fact, this type of public notification might create bad publicity and disapproval for non-compliant operators. We believe that the best way to set this process in motion is to involve third parties that can further pressure operators to adopt anti-spoofing measures. To this aim, we will try to engage the community of network operators and national CERTs. The community of network operators, especially those operators already deploying SAV, should have an incentive to promote compliance with BCP38, given that the benefits are perceived by the whole community. As for CERTs, they are a natural point of contact for abuse data, which they can then forward the relevant operators. Finally, publicly disclosing information about SAV compliance will increase awareness about the problem, showing operators that compliance is worthwhile and that non-deployer operators are penalised, and maybe even attracting regulators' attention.

3.4 Conceptual framework

In the final part of this chapter, we would like to do a brief recap of what has been discussed so far, in order to set the stage for the rest of the research.

3.4.1 Formulating the research question

We concluded Chapter 2 by discussing the incentives of network operators to deploy anti-spoofing measures. In particular, we argued that, technical limitation aside, operators have little incentives to deploy anti-spoofing because of economic barriers, such as externalities and information asymmetries, which also contribute to create a climate of uncertainty and doubt. The whole situation resembles a

tragedy of the commons in which operators do not contribute to the public good due to low individual benefits.
As previous attempts to privately notify non-deployer operators has resulted in a moderate remediation, alternative strategies to prompt compliance with anti-spoofing best practices must be considered. In this regard, a disclosure policy aimed at soliciting additional incentives out of reputation concern and peer pressure appears an applicable solution.
Therefore, in Chapter 3, we have presented the benefits of sharing security information and reviewed notification studies, funnelling our way towards the use of reputation and increased pressure as incentives. In particular, we have explored how releasing peer information can induce social comparison, and we discussed the effects of public disclosure policies on organisations.

All in all, public notifications appear fitting the case of IP spoofing because disclosing information about SAV deployment may:

- Stimulate upward comparison, and thus create motivations to deploy;
- Induce negative publicity and shame that can increase operators' incentives;
- Involve third parties who can generate pressure on operators to deploy;
- Contribute to establish that IP spoofing is not a tolerable practice and increase transparency.

The objective of this research is (as formulated in Chapter 1):

to study the effect of private and public notifications on the compliance with anti-spoofing best practices.

To release information about which operators lack anti-spoofing measures, we design Infospoofing.com, a website on which we aggregate and display the results of measurements of compliance collected by CAIDA's Spoofer Project (more on the design in Chapter 4). An email notification will be sent to non-compliant operators, in order to test the effect of privately releasing peer information. In addition, a group of operators will also be publicly notified, meaning that our website will be shared to third parties to further create peer pressure on non-compliant operators. As the number of networks found without anti-spoofing filters is a small fraction of the total number of networks tested, the wording of the notification will be crafted to emphasize this disproportion, and thus to nudge operators to comply.

Therefore, we formulate the following research question:

- *To what extent do notifications incentivise compliance with anti-spoofing best practices?*

To address this main question, additional sub-questions need to be answered:

1. *What is the effect of privately notifying non-compliant operators?*
   In light of what has been discussed in Section 3.3, we suspect that releasing information about SAV deployment can have a positive impact on the chance of remediation. In fact, according to social comparison theory, such information can stimulate upward comparison, therefore creating incentives to deploy anti-spoofing.

In order to stimulate this process, it is critical to provide operators with a meaningful term of comparison, to which they can relate. Thus, we decide to focus on a country level, releasing information in specific countries. Albeit operators within a country might may differ significantly (i.e. in size, or type of services provided), they are immersed in the same socio-cultural context, they are subjected to the same norms and they play on the same market, factors that should ensure the meaningfulness of comparisons.
To test the effect of disclosing peer information on the deployment of filters, a field experiment is designed (more in Chapter 4).

2. *What is the effect of publicly notifying non-compliant operators?*
Secondly, we seek to understand whether publicly disclosing information of SAV deployment might create additional incentive out of reputation concern and increased peer pressure. For this reason, beside notifying non-deployer operators, we aim at engaging the community of network operators (especially those already compliant who, on paper, should have an incentive to get other networks to deploy), CERTs (who are in the position to request remediating actions against abuses), and security blogs (to further increase the visibility of our website).
As in the previous question, whether third parties can generate pressure on operators and whether such pressure has an effect on the likelihood of remediation will be tested with the field experiment.

3. *Can we identify characteristics of network operators that explain differences in remediation?*
Next, we seek to understand whether operators' characteristics, such as the type of service provided, the size, and its country might predict the likelihood of remediation. Therefore, we will analyse the results of the experiment looking for correlation between operators' characteristics and remediation. Understanding which characteristics are associated with remediation can open the doors to more effective interventions in the future, as well as it can show researchers and policy makers on which type of operators to focus.

4. *What practical recommendations can be formulated on the base of the previous findings?*
Finally, we shall discuss the consequences of our finding in terms of vulnerability notification and in terms of practical policy solutions to further incentivise SAV adoption.

3.4.2 Conceptual model

On the base of what has been discussed so far, a conceptual model to depict the construct under investigation can be drawn (Figure 7).

The model consists of five constructs:

- *Information disclosure*
The first independent variable. It represents whether an operator is notified or not.

- *Publicity effect*
Publicity effect refers to the effect of sharing compliance information with third parties on the effectiveness of the notification. It is the second independent variable, which may act as a moderator on the effect of information disclosure.

- *Organisational and socio-technical factors*
  Our last independent variable refers to the characteristics of the operators and of the type network managed. Empirically, this construct is measured looking at the following variables:

  - *Organisational factors*

    - *Type of network*
      The type of the service provided by operators might contribute to shape their incentives (Internet Society, 2015). To group operators by type, we use the *AS Classification* dataset, online available on CAIDA's website[13], dataset groups ASes in three categories:
        - ◊ *Transit/access*: ASes that are either transit and/or access providers;
        - ◊ *Content*: content hosting and distribution system;
        - ◊ *Enterprise*: various organisations (e.g. university, business users) operating at the edge of the Internet that are mostly users, rather than providers.

    - *Size of the network*
      We characterise ASes also by their size, as larger and smaller operators might perceive slightly different incentives, as discussed in Section 2.2.4. We measure ASes' size by looking at the address space they announce. We use *Pyasn*, a free Python module, to lookup ASNs and retrieve the number of prefixes announced.

  - *Socio-technical factors*
    As previous research showed that the geographical location of an AS contributes to explain the lifetime of infection (Gañán, Cetin & van Eeten, 2015), we seek to understand whether the socio-technical characteristics of a country mediate the effect of disclosure. In particular, we consider the following four variables:

    - *Native English speaker*
      Notifications sent by CAIDA show that remediation is slightly more likely in English speaking countries (Claffy, 2017).

    - *GDP per capita.*
      The Gross Domestic Product (GDP) of a country represents the market value of all final goods and services produced in a nation. To compare countries' economic performances and quality of living, the GDP per capita is often used (i.e. GDP divided by the population of a country). Data about the GDP per capita of the countries selected is retrieved via the website of the World Bank[14], and refer to 2017.
      As the GDP represent a monetary value, when divided by the number of people it becomes a rate of available GDP per person, express in currency (often US$).

    - *ICT Development Index*

[13] https://www.caida.org/data/as-classification/
[14] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

The ICT Development Index is an indicator computed by the United Nations International Telecommunication Union to benchmark the level of ICT development in countries across the world (ITU, 2017b). The index is composed by 11 indicators, grouped in three major areas (ICT infrastructure and access, ICT usage and intensity, ICT skills and impact). On the base of these indicators, countries are assigned a index score, which is then used to create a ranking system. In our case, we use the IDI score of countries, and not their position in the rankings. The score represents a "grade" for the country weighted on the various indicators, and has a maximum of 1000.

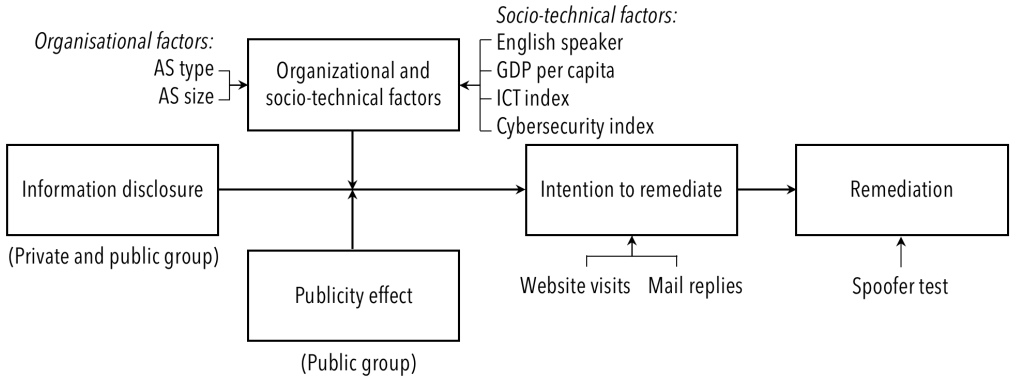
- *Global Cybersecurity Index*
  The Global Cybersecurity Index (GCI), also developed by the United Nations International Telecommunication Union, is used to measure countries' commitment to cybersecurity (ITU, 2017a). This index is composed by 25 indicators, grouped in 4 areas (Legal Measures, Technical Measures, Organizational Measures, Capacity Building and Cooperation). Similarly to the IDI, a GCI score is computed, and used to create rankings of countries. Again, we shall use the GCI score, a number between 1 and 1000, to match countries.

- *Intention to remediate*
  The intention to remediate is the first outcome we want to observe, and it represents the operators' reaction to our notifications. It is measured by looking at the two following variables:

  - *Website visit*
    The visits to our website displaying information about compliance is a significant indicator for assessing the success of the notifications. Keeping track of which operators opened the link included in the notification can provide important information about deliverability (i.e. was the mail successfully received), and operators' intention to remediate (i.e. who visits the website might be more incline to remediate).

  - *Mail replies*
    Operators' replies to our notification represent a good proxy to measure their intention to remediate.

Figure 7. Conceptual model.



- *Remediation*
  Finally, our last dependent variable is the actual remediation, that is the deployment of anti-spoofing filters. To measure remediation, we will rely on the test application of the Spoofer Project. In the notifications and on our website, we will instruct non-deployer operators to prove the correct deployment of filters by running the test.

Investigating the relations among these variables, understanding which relations hold and with which strength is necessary to establish the effectiveness of private and public notifications. Thus, we formulate these relations as our hypothesis:

Hypothesis 1:
Operators privately notified are more likely to remediate than operators not notified.

Hypothesis 2:
Operators publicly notified are more likely to remediate than operators not notified.

Hypothesis 3:
The likelihood of remediation varies significantly between operators privately and publicly notified.

Hypothesis 4:
Organisational and socio-technical factors affect significantly the effectiveness of notifications.

Hypothesis 5:
Operators who visit our website are more likely to remediate than those who do not.

3.5 Conclusions

After having discussed the problem of IP spoofing and the incentives of network operators to deploy anti-spoofing measures in Chapter 2, this chapter focused on the design of a framework to publicly notify the operators whose networks do not properly deploy anti-spoofing filters.

We began this chapter by showing that gathering, analysing and sharing security information is an form of cooperation highly beneficial for Internet security. Next, we focused on a particular form of information sharing: vulnerability notifications. After having reviewed the empirical literature on vulnerability notification and abuse reporting, we turned our attention on the public disclosure of information as a strategy to instigate compliance with norms and regulations. In particular, we saw that the disclose of peer information has been used to address problems related to the private provision of public goods (e.g. charity given and contributions to online communities), affected by social loathing. We also provided an insight on the effects of disclosure policies on corporate behaviour from a policy-making point of view. Lastly, we briefly reviewed the objectives of this research. We formulated research questions and hypothesis, and conceptualised the relations among the variable of interest in a conceptual model.

Chapter 4:
Research methodology

In the previous chapter, the research questions have been formulated, the most important variables identified and an empirical model linking these variables together has been elaborated, in order to explain the effect of notification on deployment of anti-spoofing measures.
But how are we going to test whether the relations we expect indeed hold? In this chapter, we discuss the design of the experiment to test whether information disclosure incentivise operators to deploy anti-spoofing filters. To begin with, we provide a quick overview on the experimental design, in order to set the stage for the rest of the chapter. Next, we shall describe the dataset of measurement collected by the Spoofer Project, which lies at the base of our notifications. Secondly, we present the website we design to disclose information about which network are not deploying anti-spoofing measures. Finally, we discuss in more details the setup of the notification experiment.

4.1 Overview on the experimental design

In order to evaluate the effectiveness of private and public notifications, a field quasi-experiment is designed. In particular, the design of our experiment is partially inspired by two previous studies. The first one is the field quasi-experiment by Tang, Linden, Quarterman & Whinston (2013), while the second one is the randomised field experiment by He, Lee, Han & Whinston (2016). Both these studies belong to the Spamranking Project, which, as you recall from Section 3.2, studies the effect of public rankings on organisations' outbound spam. In practice, the Spamranking Project releases rankings of the worst autonomous systems (ASes) in terms of outbound spam, which is considered a proxy for low *"network hygiene"* (Linden, Quarterman, Tang & Whinston, 2012). The group of researchers then uses the rankings to notify the organisation that send the most spam, under the hypothesis that public information disclosure might generate incentives for organisation to improve their level of security via reputation and peer pressure.

Tang and colleagues design a country-level intervention, in which ASes are first grouped by country and countries are then assigned to treatment group (for which rankings are published) and a control group. In the design of our experiment, we follow a similar procedure to assign spoofable ASes to the experimental group. Therefore, our intervention works are two levels: the measurement units of analysis (i.e. where we evaluate remediation) are ASes, while the treatment units of analysis (i.e. to measure and compare the effectiveness of the treatments) are countries, which are clusters of ASes.

As for the treatment groups, we take inspiration from the experiment by He, Lee, Han & Whinston, in which ASes are randomly assigned to three treatments: a control group (no disclosure), a private treatment group (email notification with the link to the website displaying the rankings, which is not publicly advertised) and a public group (same email notification as before, but the website showing the

rankings is publicly searchable and advertised). In this setting, the differences between remediation in the private and public groups are solely due to *"publicity effect"* (He et al., 2016), that is that the website is publicly searchable and is it advertised and promoted.

To sum up, we are going to design a field quasi-experiment in which ASes found without anti-spoofing filters are first grouped by country, and countries are assigned to three experimental conditions: a control group and two treatment groups, to distinguish the effect of public and private notification.

4.2 Data gathering and aggregation

4.2.1 List of spoofable ASes

Our primary source of data is CAIDA's Spoofer Project, which collects measurements of compliance with anti-spoofing filters thanks to volunteers that download and run a test application. In Section 2.2.4 we have introduced the Spoofer Project, briefly describing how the test application works, and in Appendix 1 we show how results are recorded and displayed on CAIDA's website. Moreover, in Appendix 1 we also discuss the process of interpretation of the results (i.e. how to generalise measurements collected at an IP address level to the level of the entire AS), providing some examples and explaining the problems that might arise during this process.

The test application works to an IP address level. However, for security concern, the precise IP address are anonymised, and the test are reported at a /24 IP prefix level. As a first level of aggregation, the spoofable IP prefixes are matched to their relative AS. Then, each AS are matched to the network operator that controls it. Large operators often manage different ASes, in order to separate different subnets and to implement different routing policies on each (for geographical reasons or for different type of end users). On the other hand, an AS is administered by a single operator (with seldom exceptions like AS2914, which is partially run by NTT/America and NTT/Asia) (Roughan, Willinger, Maennel, Perouli & Bush, 2011). Therefore, also within the same operator, ASes are relatively independent from each other (Tang et al, 2013).
For this reason we will use ASes as a measurement unit of analysis. In other words, ASes are the unit of analysis at which we evaluate the lack of filters and remediation: we group spoofable IP prefixes by ASes, and notification about the lack of anti-spoofing will be sent to the operators of these ASes.

To strengthen the effect of social comparison and reputation effect, our intervention is designed at a country level. In fact, operators in the same country are immersed in the same regulatory and market context, which, as we seen in Section 2.2.4, may play an important role in shaping operators incentives. Thus, ASes are grouped by countries, and countries are assigned to the experimental conditions as cluster of ASes/operators. Hence, countries are the unit of analysis for the effectiveness of the treatment. Information aggregated to a country level is in fact easier to manage, and can be passed to actors that operate at a national level to further promote our initiative.

In practice, this aggregation process is done in different steps. First, test results are aggregated to the /24 IP prefix (i.e. a block of 256 adjacent IP addresses). A significant problem we need to deal with is

that, despite a single test assumes one of four possible outcomes (*Spoofable, Blocked, Rewritten* or *unknown*, as explained in Section 2.2.4), when looking at the sequence of test collected from the same IP prefix over time we notice that results are not always consistent. As we discuss in Appendix 1, the sequence of results from the same prefix might include a series of "blocked" test followed by a "spoofable" test. For this reason, we formulate the following metrics to define the status of an IP prefix given the tests on its IP addresses:

- *Spoofable*: an IP prefix from which the most recent test shows evidences of spoofing;
- *Mixed*: an IP prefix from which at least one test shows evidences of spoofing, but the most recent test does not;
- *NAT-blocking*: an IP prefix from which the majority of test is "rewritten" (with no evidences of spoofing);
- *UnSpoofable*: a prefix from which the majority of test is "blocked" (with no evidences of spoofing).

Note that the distinction between *Spoofable* and *Mixed* reflects the difficulties in clearly categorise a prefix. In fact, as we show in Appendix 1, it might be not clear whether sequences of *spoofable-unspoofable* results happens because the operator remediates or because the test is performed on close, but different segments of the network, implementing different routing policies. Thus, we felt the need to denote these situations of uncertainty as *mixed*, to leave operators the benefit of the doubt in the absence of consistent evidences of spoofing.

Next, we extend these metrics from an IP prefix level to the entire AS:

- *Spoofable*: an AS with at least one "spoofable" IP prefix;
- *Mixed*: an AS with more "mixed" IP prefixes than "spoofable" IP prefixes;
- *NAT-blocking*: an AS with a majority of "NAT-blocking" IP prefixes (and no "spoofable" prefixes);
- *UnSpoofable*: an AS with a majority of "UnSpoofable" IP prefixes (and no "spoofable" prefixes).

In practice, we wrote a Java program to download all the test results collected in a country from CAIDA's website. Once we download all the test results, we apply the metrics just presented to produce a list of ASes found without anti-spoofing filters in a given country, which will be used as an input for assigning countries to the experimental groups. Next, we design a website on which we aggregate and publish information about ASes found without anti-spoofing for the countries assigned to the treatment groups. In the following subsection, we introduce the design of such website.

4.2.2 Infospoofing.com

As our goal is to disseminate information about compliance with anti-spoofing best practices, and stimulate comparison between operators, the attention to the type of information disclosed is critical. On the one hand, information released should be complete and precise, on the other hand, it should have specific focus to enable comparison, thus avoiding risk of information overload.

As we show in Appendix 1, the webpage of the Spoofer Project already provides a good interface to browse measurements results and to monitor networks' deployment of ingress filtering. In addition to the reporting engine, CAIDA's website has a page of general statistics, in which the measurements are summarised and graphed. However, these statistics are only displayed at a global level, without the possibility to drill down and obtain specific information about countries, ASes and prefixes. Moreover, CAIDA's website is addressed to a relative technical audience, and it is easy to get lost in the high level of detail provided.

To better engage operators and to stimulate comparisons, we believe that we first need to better aggregate CAIDA's measurements, selecting specific pieces of information that can be relevant, and avoiding providing excessive details which can undermine our aim of stimulating comparisons. Therefore, we design our website, named Infospoofing.com, on which we aggregate and display CAIDA's measurements with a country-specific focus. For each country in the treatment group, we show statistics on the tests results both at a IP prefix level and AS level in form of pie charts. In addition, we explicitly show the ASes found without anti-spoofing filters on a histogram, whose height represents the number of spoofable prefixes (shown in Figure 9). In this way, we create a sort of ranking system of the worst ASes. Note that we treat all ASes in the same way, without distinguishing between type or size of the ASes, and with no distinction between IPv4 and IPv6 interfaces.

As we saw in Section 3.3.2, exposing the name of the non-compliant networks is functional to provoke bad publicity and public disapproval. However, to stimulate comparisons, the information about which operators already deploy anti-spoofing plays a fundamental role, as they would be the term of comparison for not deployer operators. Looking at the dataset, it emerges that that the networks lacking anti-spoofing are a small fraction of the total number of networks tested. Since highlighting this disproportion can contribute to our goal, the language on the website is crafted accordingly. For instance, the histogram is presented with the sentence: "*Only 9 ASes were found without anti-spoofing filters (10%), is yours one of these?*"

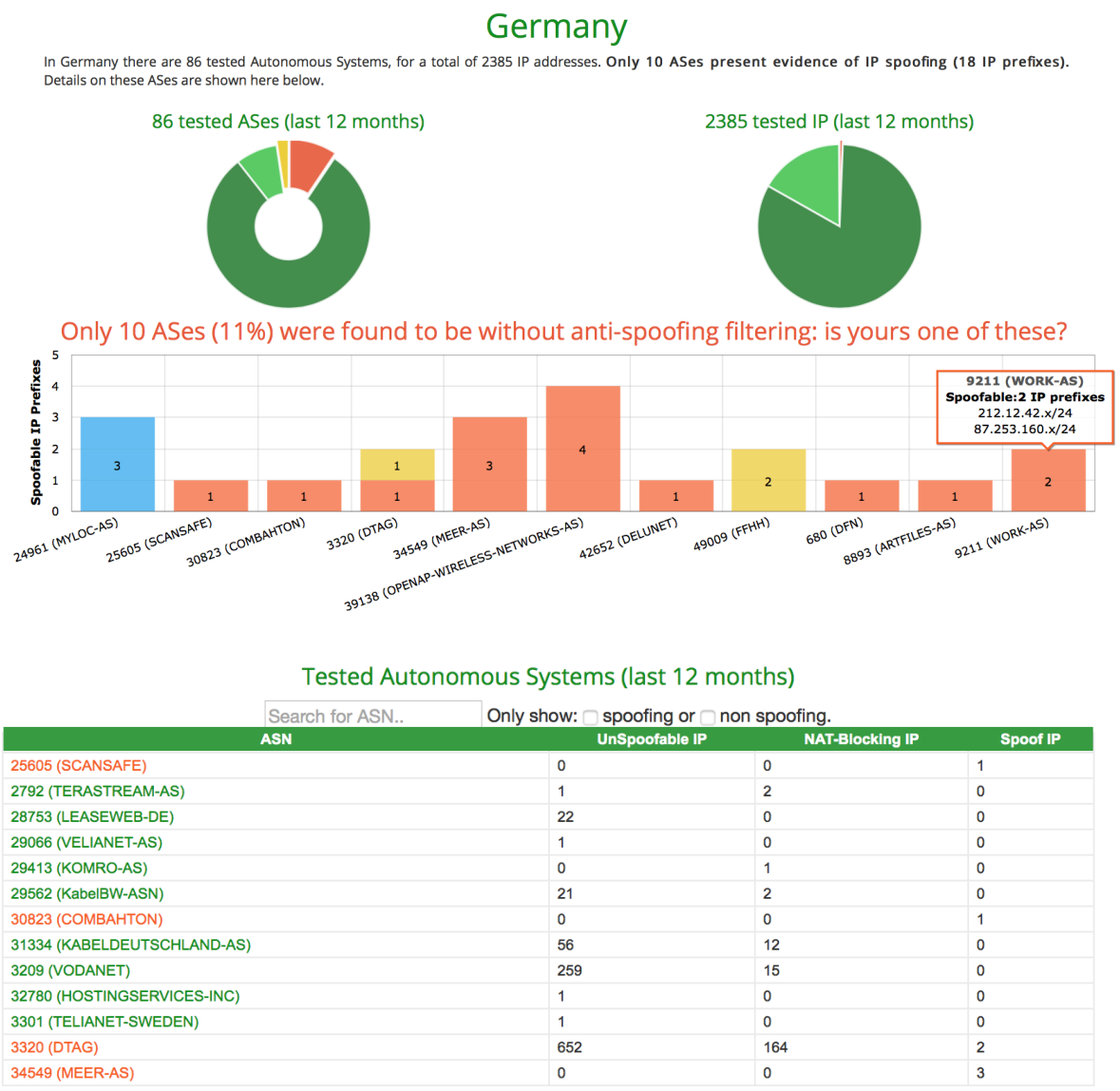
Moreover, we design a table to show the results of all the tested ASes. For each AS, we show the ASN and the AS name (in red if spoofable, in green if not) and the number of tested prefixes grouped by status. For each AS, a link to CAIDA's website is included, in order to provide additional details on the measurements. The table also includes some filtering mechanisms to show only compliant or spoofable ASes, a function to search by AS name and number, and the possibility for users to automatically lookup their AS. In addition to the specific countries pages, we also created a homepage with an interactive map that show countries details and that allows users to explore further statistics. Moreover, additional graphs to compare results among countries are designed. Finally, we also create a "remediation" page, on which we provide information about the best practices to prevent IP spoofing. We include links to the most relevant publication and to configuration examples, to the MANRS initiative and to the Spoofer Project. Moreover, we instruct spoofable operators to contact us to communicate their intention to remediate, so that we can change the status of their AS. In particular, when an operator contacts us, we will change the status of the AS from *spoofable* (displayed in red) to *remediating* (displayed in blue), to show that the operator is taking care of the problem. We also ask

operators to download the Spoofer application and to take the test to prove that anti-spoofing filters have been deployed.

Presenting precise and accurate information is critical for our intervention. If we display an AS on our "spoofing list", we must be sure to have enough evidences about its lack of filtering. However, as mentioned in the previous section, this is not always easy. For example, positive tests can be fairly old, and operators might have since remediated without doing additional tests. The presence of a high number of similar false positives might easily undermine the credibility of our website. Therefore, we seek to reduce the number of false positives by manually checked every spoofable IP before publishing it. In particular, we evaluate the general situation of the AS: if a prefix is marked as spoofable, but other spoofable prefixes in the AS have been remediated in the meantime, we discard that prefix. For the same reason, we consider only measurements collected from 2017 onwards. Finally, as explained in the previous section, we add the status *mixed* (displayed in yellow), to refer to IP prefixes from which at least one test showed evidences of spoofing, but the most recent test does not.
Unfortunately, this process of analysis of borderline cases is also prone to the risk of creating false negatives, spoofable prefixes that are incorrectly discarded. However, we decided to tolerate this risk, in order to avoid the potential negative impact of false positives on the success of our intervention.

Figure 8. Homepage of our website.

Figure 9. Website page showing compliance information in Germany



ASN	UnSpoofable IP	NAT-Blocking IP	Spoof IP
25605 (SCANSAFE)	0	0	1
2792 (TERASTREAM-AS)	1	2	0
28753 (LEASEWEB-DE)	22	0	0
29066 (VELIANET-AS)	1	0	0
29413 (KOMRO-AS)	0	1	0
29562 (KabelBW-ASN)	21	2	0
30823 (COMBAHTON)	0	0	1
31334 (KABELDEUTSCHLAND-AS)	56	12	0
3209 (VODANET)	259	15	0
32780 (HOSTINGSERVICES-INC)	1	0	0
3301 (TELIANET-SWEDEN)	1	0	0
3320 (DTAG)	652	164	2
34549 (MEER-AS)	0	0	3

4.3 Pre-test and post-test measurements

As you recall from Section 2.2.4, a promising approach to increase the coverage of the measurements of compliance collected by the Spoofer Project is to use crowdsourcing marketplace to recruit volunteers to download and run the test application, in exchange of small monetary compensations.

In our experiment, we use crowdsourcing to enable pre-test and post-test measurements, in order to verify if operators have indeed deployed filters. Once countries are assigned to the experimental groups,

we will launch country-specific pre-test crowdsourcing campaigns, in order to gather new measurement. In particular, the use of pre-test measurements might identify new ASes to include in our sample of spoofable ASes. At the same time, new measurements might cover ASes already tested in the past, reducing the number cases in which the measurements are too old.
After the experiment is conducted, we will launch a second round of crowdsourcing measurements (targeting only the spoofable ASes we notify) to assess whether remediation occurred.

One important consequence of the use of crowdsourcing platforms is that the selection of the countries to include in the experiment is constrained by the geographical coverage of the crowdsourcing platform. In other words, we first need to select the crowdsourcing platform, and, on the base of its demographic, we can select countries to include in the experiment. This is because it would be pointless to include countries in which there are no users available to download the test application.

As mentioned in Section 2.2.4, previous studies investigated the use of crowdsourcing marketplaces for measurements of compliance with anti-spoofing best practices. In particular, Lone et al. (in press) surveyed six different crowdsourcing platforms, and found that 51% of the submission came from Prolific Academic, a British platform that also offered the largest geographical coverage. Therefore, we will focus on Prolific Academic to recruit participants to download and run the Spoofer test.

4.4 Country selection and assignment

4.4.1 Country selection

True experimental design requires random assignment of participants to the treatment groups. Despite desirable in ideal environments, complete randomisation can be problematic in our context. In fact, we suspect that there might be additional variables that can affect the effectiveness of the treatments.

First, as our website and the notifications are in English, we assume that the English-speaker countries will respond differently to our intervention that non-English-speaker ones (as also noticed by researchers of CAIDA (Claffy, 2017)). We also expect that the effectiveness of our intervention is conditional to socio-technical factors such as the economical posture of countries, the level of activities of cybersecurity institutions, development of the ICT infrastructure, and the extension of the problem.
Thus, to ensure a selection of countries that is comparable in regard to these factors, countries are preliminary grouped on the base of GDP per capita, Global Cybersecurity Index, ICT Development Index and number of spoofable IP prefixes. A cluster analysis provides the perfect tool for this operation.

Cluster analysis refers to the process of grouping items so that the resulting groups present high within-group homogeneity (with respect to some independent variable), and high between-group heterogeneity. Once clusters of similar countries are computed, a subset of countries from each cluster can be assigned to the experimental conditions. Again, in spite of the desirability of random assignment, from each cluster we deliberately select the three countries (one for each treatment) with the highest number of spoofable IP prefixes. In fact, given the relatively small number of spoofable networks identified, we prioritise having a sample with the highest possible number of network to notify.

As mentioned in the previous subsection, we need to first identify the countries in which the crowdsourcing platforms is active, and then assign these countries to the experimental groups. Thus, we need to investigate the demographic of Prolific, in order to identify the countries with the higher number of users.

When creating a new study, Prolific provides the possibility to select some screening criteria on participants (e.g. participants' country of residence). The platform then provides details on the number of users that fit the screening criteria. Thus, we created a fictional study, setting as screening criteria the country of residence of users, and selected one by one all the possible countries.
In this way, we created a list of the countries most covered by the platform (with at least 80 users). This list is presented in Table 2 below, which also include the score of each country on the independent variables used to group countries together.

Table 2. List of countries most covered by Prolific.

Country	*N. users*	*Matching variables*			
		N. spoofable IP prefixes	ICT Development Index (score)	Global Cybersecurity Index (score)	GDP per capita (million US $)
United Kingdom	16745	53	865	783	40 341
United States	8199	376	818	919	57 638
Canada	603	34	777	818	42 158
Portugal	556	1	713	508	19 840
Germany	475	14	830	679	42 070
Italy	312	6	704	626	30 675
Spain	263	1	779	718	26 640
Australia	249	32	824	824	49 928
Netherlands	181	11	849	760	45 670
Ireland	154	1	802	675	63 826
France	141	8	824	819	36 855
Greece	130	1	723	475	17 930
Poland	115	23	689	622	12 421
Belgium	94	1	781	671	41 158
Turkey	85	18	608	581	10 862
Sweden	82	6	841	733	51 943

4.4.2 Cluster analysis

The list of countries presented in Table 2 is used as input for a cluster analysis, with the goal of creating group of countries that are similar with respect to the number of spoofable prefixes, ICT Development Index, Global Cybersecurity Index and GDP per capita[15].
We choose to perform an Agglomerative Hierarchical Clustering, which takes a bottom-up approach to data clustering. The algorithm begins by considering all the items as individual clusters. Then, the distance between all couple of items is computed, and the closest items are merged together. This step of computing the distance between items (or clusters) and merging the closest is iterated until only a macro cluster remains. Two important factors in the algorithm are the metric used to assess the distance between clusters and the linkage criterion to merge them. Among the various metrics to measure the distance between items, we opted for using the Euclidean distance. As for the linkage function we use Ward's method, which minimise the total variance within-cluster. In other words, each iteration merges two clusters with minimum between-cluster distance.

The outcome of the clustering process is represented in a dendrogram, a tree structure that depicts that hierarchy of clusters, showing the merges done at each iteration. The height of the dendrogram represents the (Euclidean) distance between cluster. Figure 10 (left) shows the dendrogram for our data, which seems to identify three clusters. In fact, the dendrograms reveals that three cluster are created in the first iterations, within a distance of slightly larger than *0.5*. Furthermore, the dendrogram shown no additional merges until approaching distance of *2*.
In order to show that the optimal number of clusters is indeed three, we follow the *elbow method* and we further analyse the total intra-cluster variation (a.k.a. within cluster sum of squared errors). Figure 10 (right) shows the within cluster variation as a function of the number of cluster produced. We see that the *elbow* (i.e. the point in which additional clusters provide smaller reductions in sum of squared errors) is on $k=3$ clusters. The R code for the country assignment is reported in Appendix 3, together with an overview of the descriptive statistics of the dataset.

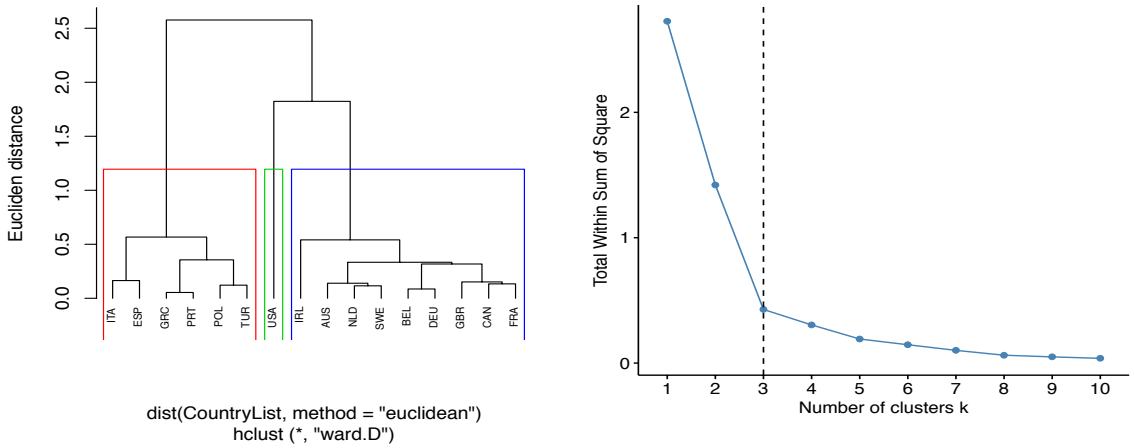
All in all, we started by considering the 16 countries in Table 2, and we have clustered on the base of the ICT Development Index, Global Cybersecurity Index and GDP per capita, producing 3 clusters of countries, as highlighted in the dendrogram in Figure 10. It is interesting to notice that the United States form a cluster on their own, probably because the high number of spoofable IP prefixes and the high score on the other matching variables make the US stand out as an outlier.

4.4.3 Country assignment

The next step is to assign countries from each cluster to the experimental conditions. Besides the cluster containing only the United States, the other clusters count more than three countries. As we have three treatment groups (control group, private group and public group), from each cluster we select triplets of countries to assign to the treatments, so that countries in the same triplet are mutually comparable. In particular, we select the countries with the highest number of spoofable IP prefixes. For example, from the first cluster, one triplet is Italy, Poland and Turkey. Unfortunately, the second possible triplet

[15] Note that these are the *socio-technical factors* mentioned in Section. 3.4

Figure 10. Dendrogram (left) and elbow method (right).



from the first cluster (Spain, Greece and Portugal) contains only 3 spoofable prefixes in total, so we decide to drop this triplet. From the other cluster, instead, we select Australia, Netherlands, Germany, the United Kingdom Canada and France. This second group of countries can be further divided to create a triplet of native English speaker countries (Australia, the United Kingdom and Canada) and another triplet of non-English speakers (France, Germany and the Netherlands). This distinction might come handy because both the notifications and the website to disclose compliance information are in English. Moreover, researchers of CAIDA observed that remediation appears to be slightly more likely to occur in English speaker countries (Claffy, 2017).

To sum up, starting from the 16 initial countries we have created three triplets of countries coming from the same clusters, and thus comparable. Figure 11 shows the similarities between the triplets of countries selected in form of a histogram (after normalisation of the score on the matching variables).

The final step is to assign countries from each triplet to the experimental groups. We deliberately decide to assign Italy, the Netherlands and the United Kingdom to the public group, in order to facilitate the promotion of the website in these countries. The rest of the countries are randomly assigned to control group and private group. As notifications are sent to the operators of the ASes, we group the spoofable IP prefixes by AS. Finally, we manually check the state of each AS, in order to discard false positive, as explained in Section 4.2.2. The outcome of the assignment is reported in Table 3, together with the final number of ASes included in our experiment (before the pre-test crowdsourcing measurements).

Figure 11. Visualization of the countries selected for the experiment grouped by cluster.

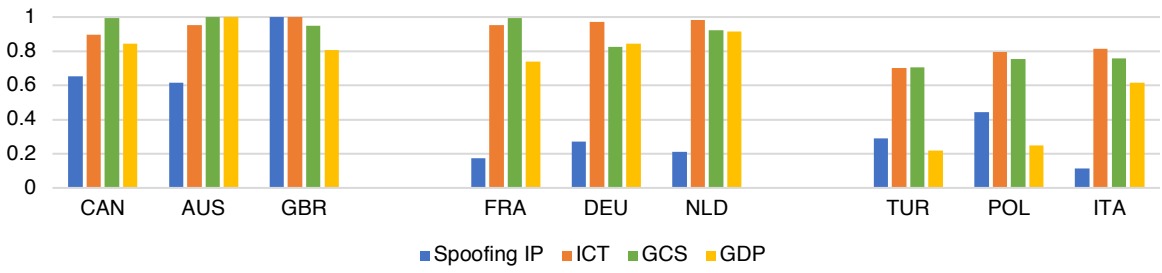


Table 3. Experimental groups

	Treatment					
	Control group		*Private disclosure*		*Public disclosure*	
	Country	Spoof AS	Country	Spoof AS	Country	Spoof AS
Triplet 1	Canada	17	Australia	13	United Kingdom	20
Triplet 2	France	4	Germany	10	Netherlands	9
Triplet 3	Turkey	7	Poland	4	Italy	6

4.5 Pre-test crowdsourcing measurements

Once the experimental groups are formed, we can proceed with the pre-test crowdsourcing measurements of compliance with anti-spoofing. The aim of these measurements is to both obtain fresh tests on networks already tested and to test new networks. Moreover, if we find any spoofable networks, there is the possibility to test again these networks after the interventions, so that we can evaluate if the operators have remediated.

Using crowdsourcing platforms to measure compliance with anti-spoofing filters is a non-trivial task, and many complications have arisen during the design phase. The main problem is that crowdsourcing platforms are envisioned for so-called "human-intelligence tasks" and not for conducting measurements of network properties. This forced us to create an ad-hoc measurement infrastructure to facilitate the data collection and to manage participants' submission. We discuss in more details the design the crowdsourcing study in Appendix 4, while here we report the results of the pre-test measurements of compliance conducted in the 9 countries selected in the previous section.

4.5.1 Pre-test results

The pre-test crowdsourcing measurements started in mid-January, and lasted for three weeks. We rewarded participants that successfully submitted test results with £1, and offered a partial compensation of £0.10 to participants that took part in the study but did not manage to submit the results. In total, we spent around £350 (including participants compensation and platform fees) to collect 202 valid tests from 110 different ASes. Of these 202 tests, 12 revealed networks lacking anti-spoofing filters (with a partial overlap with spoofable networks previously tested by CAIDA). Interestingly, we tested 15 new ASes that were not previously tested by CAIDA.

Table 4 shows the final number of ASes that have shown consistence[16] evidence of spoofing from January 2017. In the brackets we report the number of spoofable ASes identified by the pre-test measurements.

[16] Note that we excluded some ASes for which evidence of spoofing were not consistent, as explained in Section 4.2.2

Table 4. Final number of ASes included in the experiment.

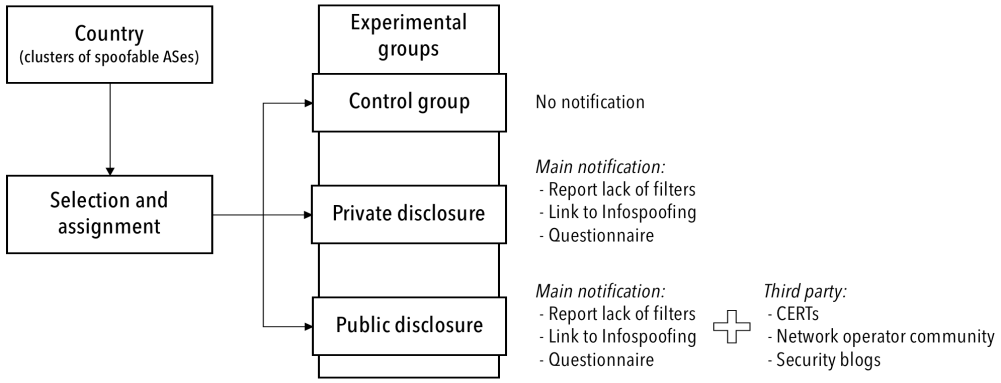
	Treatment					
	Control group		*Private disclosure*		*Public disclosure*	
	Country	Spoof AS	Country	Spoof AS	Country	Spoof AS
Triplet 1	Canada	19 (3)	Australia	15 (3)	United Kingdom	22 (2)
Triplet 2	France	5 (1)	Germany	11 (1)	Netherlands	9 (0)
Triplet 3	Turkey	7 (0)	Poland	4 (0)	Italy	7 (2)

4.6 Treatments

4.6.1 Experimental groups

We form three experimental groups to test the effectiveness of private and public notification against a control group. The first experimental group receive no treatment (i.e. control group). Operators in the second group are notified with an email to report the lack of anti-spoofing on their network, providing additional information on the number of networks found with and without filters, a link to our website, and a short questionnaire (we shall call this treatment "private disclosure group": peer information is privately shared only to involved operators). Lastly, operators in the third group receive a similar email notification of the second group. Additionally, for countries assigned to this third group we promote and advertise our website, sharing it with a group of third parties that can act as intermediaries to pressure on non-compliant operators (we shall call this treatment "public disclosure group": peer information is publicly revealed to operators and a selection of third parties). Differently from the private group, the mail notification for operators in the public group explicit mentions that our website is publicly advertised, and that third parties have been informed. Figure 12 shows and overview of the design of the experiment and the structure of the treatments.

Figure 12. Experimental groups



4.6.2 Notifications to non-compliant operators

Operators of spoofable ASes in the two treated groups are notified about the lack of anti-spoofing filters via email. An example of the notifications is reported in Appendix 5. We also designed a short questionnaire to include in the notification, reported in Appendix 6.
In the notification, we include information about the results of measurements in the operators' country. The message is crafted to highlight that networks found without anti-spoofing are a small percentage of all tested networks. In other words, we attempt to nudge operators to deploy ingress filtering, by proposing that compliance is rewarded (i.e. positive reinforcement) and that lack of compliance can result in shame and bad publicity (i.e. negative reinforcement).
Beside formulating the body of the message, sending vulnerability notifications involves other important decisions, chief among which the choice of the address to which send the message. Previous research has focused on the use of abuse contacts retrieved via WHOIS lookup (Cetin, Ganán et al., 2016; Li, Durumeric et al., 2016). As we mentioned, WHOIS is a standards query/response protocol, whose aim is to provide information about owner (or in general the entity responsible) of a determinate Internet resource (Daigle, 2004). In particular, the WHOIS record for an AS should include an "abuse contact" field, which indicates the email to which report abuse from that AS. In our experiment, operators of spoofable ASes are notified to this address.

4.6.3 Notifications to third parties

For countries in the public disclosure group, information about which networks are found without anti-spoofing filters is shared with a selection of third parties. In particular, our intention is to engage actors that can further incentivise the adoption of anti-spoofing by creating pressure on non-deployer operators. To be more precise, we would like to engage the community of network operators, national CERTs and that security bloggers.

Network Operators Groups (NOGs)

All network operators have incentives to reduce the risk of DDoS attacks. In particular, those operators that have already deploy anti-spoofing (arguably those most concerned with security) are in the position to create peer pressure on non-deployer operators. This group of deployer operators has also an economical incentive to act as third party: they have already faced the costs of security and of implementation of filtering. Having given their contribution to Internet security, it is reasonable to assume they want other operators to do the same. Moreover, we want to generate awareness about our website, so that it can work as a deterrent for operators that have not been tested yet.
Network operators' role is both to contact peer non-compliant operators and at the same time to further disseminate the information. One way to reach network operators is through Network Operators Groups (NOGs), regional and often informal groups of operators whose aim is to facilitate information sharing and promoting debate on matters of mutual interest, mainly via mailing lists and conferences. Thus, we send an email to the NOGs of the United Kingdom, the Netherlands and Italy, reporting the results of compliance measurements in the respective countries and including a link to the relative page on our website. Moreover, we ask operators to reach out to non-compliant operators, with the aim of pressuring them to deploy anti-spoofing filters.
An example of the notification sent to NOGs is reported in Appendix 7.

Computer Emergency Response Teams (CERTs)

CERTs are groups of experts that handle security incidents on a national level (founded by governments), or on larger scale (EU-CERT). Typically, CERTs already have the technical infrastructure and the procedures to forward vulnerability information to the administrators within their authority (Stock, Pellegrino et al., 2016). The role of CERTs in our intervention is to contact non-compliant operators and to instigate them to remediate. In particular, we contact the national CERTs of the countries in the public group, asking for their collaboration to pressure on non-compliant operators.

Contacts for CERTs are retrieved on the website of the Trusted Introducer[17], an initiative established by the European CERT community which provide a list of email contacts.
Appendix 8 reports an example of notification to CERT.

Security blogs

Finally, we will contact security bloggers in order to better promote the website, for example, by posting about the amount of spoofable ASes in their country on a security blog or on the social media. Their publication can boost the dissemination of information, overcoming language barriers that would limit our study to English-speaking countries and even involving the public opinion. The risk of making headlines is definitely a deterrent for network operators to not implement filtering, and might contribute to instigate non-deployer operators to remediate.

After a quick research about the security blogs most active in the countries in the public group, we have identified the following:

- The Register (a British blog about IT and security)
- Secutiy.nl (a Dutch blog about cybersecurity)
- Securityinfo (an Italian blog about cybersecurity)

Appendix 9 shows an example of notification to a security blog.

4.7 Measuring remediation

4.7.1 Intention to remediate

As you recall from Section 3.4, our conceptual model includes *operators' intention to remediate* as a mediating construct, which is measured by looking at the replies to our mail notifications and to the visit to our website.

The replies to our notification will be treated as qualitative data, and will be also used to investigate which factors prevent operators from deploying filters. In particular, we seek to understand whether operators were aware about the problem, if they have filters deployed (perhaps misconfigured) or if they do not have filters at all. In this last case, we are interested in understanding if the lack of filters

[17] https://www.trusted-introducer.org

is due to technical problems, high costs of deployment or other reasons. Thus, when operators reply to our notification, we will follow up with few questions on the reasons of their lack of compliance.

The other variable we observe are the visits to our website. The link to the website in the notification include a unique token (i.e. an identifier that unique for each notification we send both to operators and to third parties), so that we can track the visits to our website. In this way, we are able to monitor which operator opens the website. Moreover, we can also assess the effectiveness of the dissemination of information via third parties (e.g. how many people view the website via NOGs).

Operators' visits to our website is also a good proxy for assessing the success of the notification deliverability. Moreover, we assume that operators who visit our website will be more likely to remediate.
Surely, we cannot tell whether operators who do not open the website did not received the notification or simply ignored it. However, in both cases, being able to identify these operators might be useful to profile operators with low security incentives, either because their WHOIS record is not accurate or because they ignore the notification.

4.7.2 Deployment of filters

The aim of our experiment is to test whether notifications induce operators to deploy anti-spoofing filters. As the main tool to measure deployment of filters is the Spoofer test, which requires an insider to run the test application, we instruct the operators we notify to do the test to prove that they have indeed remediated. Moreover, post-test crowdsourcing measurements will be conducted to test again the spoofable networks after the experiment. Nevertheless, the pre-test measurements identified 12 spoofable ASes, and so, for the remaining 87 ASes included in the experiment we have to rely on the tests done by operators (as instructed by our notification and website) and by volunteers (collected independently from our experiment).

The are some limitations to this approach for measuring remediation.
Firstly, as mentioned in 2.2.4, deploying anti-spoofing filters might require time, especially if it must be done manually or from scratch. In fact, we assume that there is a difference between operators that already partially deploy filters (e.g. in case of configuration errors, in which the time needed for remediating could be short), and operators who instead do not have filters at all (in which case the time needed to deploy filters might be longer than the time span of our experiment). In this last case, we might not be able to measure the deployment of filters.

Secondly, we assume that operators are able to run the Spoofer test on the precise IP prefix that showed evidences of spoofing (which we report to them). In reality, we do not know if this is always the case, leading to possible situation where operators claim remediation showing test results on an IP prefixes different to the one we reported, yet relatively close. Consider, for example, the case of an operator managing the prefix 130.251.0.0/16 (which includes 2^{16} addresses). Imagine that we report the lack of

filters on the prefix[18] 130.251.183.0/24, and that, after the notification, a new test shows the presence of anti-spoofing filters on the prefix 130.251.184.0/24 (and no new tests are collected on the original prefix we reported). Since the two prefixes mentioned above belongs to the same /23 prefix (i.e. 130.251.180.0/23), can we conclude that filters have been deployed also on the original prefix we reported? Have filters been deployed on the entire /16 prefix? What would be the ideal the prefix length on which evaluating remediation (e.g. /24, /23, ..., /16)?
These are all important questions, and different solutions leads to different interpretation of the results, thus affecting our analysis of remediation. We opted for using the strictest interpretation, referring to the /24 prefix. In other words, we will consider a prefix remediated by looking only to the tests received from that /24 prefix. Note, however, that the /24 prefix length was chosen by CAIDA for the purpose of anonymising the spoofable IP address out of security concerns, not for the aim of assessing remediation.

In practice, the potential mismatch between the spoofable prefix and the prefix on which new tests are conducted may cause additional headaches. For example, AS9105 (TISCALI-UK) showed evidences of spoofing from the following prefixes: 79.75.14.0/24, 79.75.16.0/24, 79.75.19.0/24, 79.75.29.0/24, which belong to the larger prefix 79.75.0.0/19. In similar cases of multiple spoofable prefixes from the same AS, if we observe remediation on just one of the spoofable prefixes, we will consider the whole AS remediated.

A final limitation in the process of evaluating remediation is related to the presence of NAT. Suppose, for example, we have evidence of spoofing from a given prefix. After we sent the notification to the operator, imagine that new tests are collected, but their result is *rewritten* (meaning that the test source address of the test packets are rewritten by a NAT), so that we cannot conclude about the presence of anti-spoofing filters, as explained in Section 2.2.4. In these cases[19], we will consider the AS still spoofable.

4.7.3 Defining remediation

Following from what has been discussed in this section, how can we evaluate if an operator remediates the problem by deploying anti-spoofing filters?

Given the limitations of measuring remediation with the Spoofer test just presented, and given the limited time span of our experiment, we opted to assess remediation on the base of both operators' replies to the notification and the (available) Spoofer tests.

First, we look at operators' mail replies. If their message clearly states that they have deployed filters, we will consider that AS remediated (also, note that the day in which we receive the message will be considered as the day in which remediation occurs). Additionally, we will ask the operators to run the Spoofer test to prove they indeed deployed filters.

[18] Remember that we only have data referring to /24 prefixes because of the anonymization process automatically done by CAIDA.
[19] Unlikely, this is not an unrealistic situation, and many prefixes present this problem. Check, for example, the prefix 106.69.4.0/24: https://spoofer.caida.org/recent_tests.php?subnet_include=106.69.4.0/24

In case of a mail reply claiming remediation, but in lack of new tests, we still consider the AS remediated. In fact, operators that spend time to reply already show their involvement with the problem, and we argue that they have little incentives to lie about remediation, as it would be easier for them ignoring the notification rather than pretending to have deployed filters. Moreover, as it has also emerged during the preliminary interviews, trust is a main component in the relations between operators on the Internet, so we decide to believe to operators when they say to have fixed the problem.

Secondly, we will control the results of new Spoofer tests.
If an operator does not contact us, but new tests reveal the presence of filters on an AS we notified, we will consider that AS remediated. Note that, as explained in the previous section, a prefix is considered remediated only if new tests are performed on the same /24 prefix.

In presence of both the mail reply and new test confirming the deployment of filters, the AS will be considered remediated. Note that, in this case, we accept also test results coming from IP prefixes different from the one we report as proof of remediation.

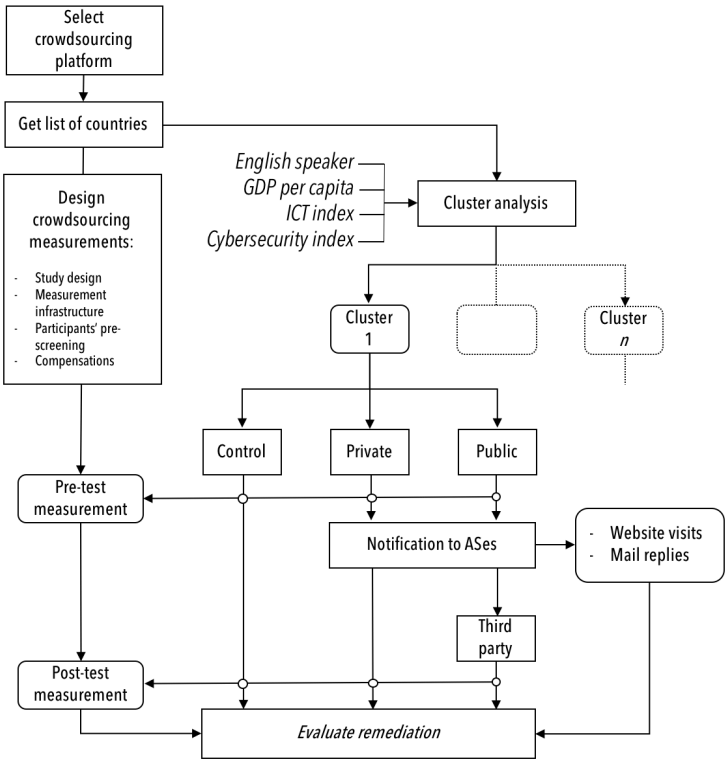
Naturally, if new tests reveal the lack of filters, the AS is considered still spoofable, despite any eventual evidences of remediation previously observed.

Finally, due to the limited duration of the experiment, remediation is assessed in two moments. First, we will assess remediation during the 25 days of the experiment, using the metrics formulated above. The analysis of remediation, and consequently of the effectiveness of the notification, is conducted on the data gathered in these 25 days.
After the conclusion of the experiment (when the notification website will be no longer updated), we will look again at eventual new Spoofer test received from the prefixes notified, to check if these new tests are coherent with the previous analysis.

4.8 Putting all together

The overall design of the whole experiment is schematised in Figure 13.
The first step is to select the crowdsourcing platofrm to conduct pre-test and post-test measurement of compliance. Based on the results of a previous experiment aimed at surveing different crowdsourcin platforms (Lone et al., in press), we select Prolific Academic due to its larger geographical coverage (compared to the other platforms previously identified).
Next, we analyse the demographics of Prolific, in order to produce a list of the countries with the highest number of users. Countries are then clustered together to create groups of countries that are similar in terms of GDP per capita, ICT Development Index, Global Cybersecurity Index and number of spoofable IP prefixes. From each resulting cluster of countries, triplets of countries are selected and assigned to the three experimental groups (control group, private notification and public notification). In particular, from each cluster we select the countries with the highest number of spoofable IP prefixes. Moreover, we manage to create a triplet of English speacker countries, so that we can control eventual effects of the language on the success of the experiment.

Figure 13. Review of the experimental design.



We deliberately assign the United Kingdom, the Netherlands and Italy to the public group, in order to facilitate the promotion of our initiative in these coutries. The rest of the countries are randomly assigned to control and private group.

Next, pre-test measurements of compliance are conducted, by using crowdsourcing to recruit volunteers to download and run the Spoofer test. During the pre-test measuemrent, 12 spoofable netwokrs have been identified.

At this stage, we can implement the pages of our website to show statistics about the adoption of anti-spoofing, but only after checking accuracy of the data (i.e. to make sure no new tests showed evidence of remediation). Note that our website reports statistics only for the countries in the private and public group, information is not shown for countries in the control group.

Once the content of the notification is written, and the contact information for spoofable ASes are retrived via WHOIS lookup, notifications are sent to the operators in the private and public group. In addition, third parties (i.e. national CERTs, NOGs and security blogs) are contacted for countries in the public group.

At the end of the experiment, a second round of crowdsourcing measuraments is conducted, to test the networks notified and to evaluate remediation. In addition, remediation is also assessed by looking at operators' replies to our notifications.

4.9 Conclusions

In this chapter we described the research method. We began by discussing how test of compliance with anti-spoofing best practices are collected and displayed by CAIDA. We introduced our own metrics, in order to extend results of tests collected to a IP address level to the entire AS and, subsequently, to the country level. We explained the design of our website, and its role in our notifications. Next, we turned our attention to the experimental design. We described the tree experiment groups and the treatments. We discussed the content of the notifications, the recipient of the notifications and their role in the experiment. Finally, we reviewed the variables we are interested to observe and the way in which we are going to measure them.

Chapter 5:
Results and data analysis

In this chapter, we discuss the results of the experiment, and we perform the analysis needed to test our hypothesis. We begin with a qualitative discussion of the reactions to our notification: we analyse the mail replies of both operators and third parties, and we provide an overview of the visits to our website. Then, we evaluate remediation by looking at the new Spoofer test collected during the period of the experiment. Finally, we perform statistical analysis to assess the effectiveness of the notification. In particular, we perform survival analysis to assess the effectiveness of the treatments, by comparing the probability of remediation in the three treatment groups. Next, we try to identify factors that can be used to predict remediation. In fact, though the country assignment already controls for socio-technical factors (economic posture, development of ICT, activity of cybersecurity institution and English proficiency), we seek to understand whether additional organisational factors such as the size and the type of the network can be associated with the likelihood of remediation.

5.1 Notification analysis

In Chapter 4, we have selected and assigned nine countries to the three experimental groups, and we have also reported the results of the pre-test crowdsourcing measurement. The first columns in Table 5 summarises the situation in terms of spoofable ASes in these nine selected countries (column *Spoof ASes*). We began sending notification on March 9th, and the experiment lasted for 25 days. In this section, we discuss the reactions to our notifications. First, we describe operators' feedback, analysing their replies and the visits to our website. Next, we discuss third party engagement.

5.1.1 Notifications to operators

We sent 66 email notification to the 68 ASes in the private group and in the public group (two couples of ASes, one in Australia and one in the United Kingdom pointed to the same abuse contact, and we aggregated these couples of ASes in the same notifications). There have been three cases of delivery failures: in two of them, we successfully sent the notification in a second attempt, in the last case, instead, the delivery failed again because the recipient mailbox was full, and only one email address was provided as abuse contact.

Table 5 reports the response rate to our notifications in the 6 countries for in the two treated groups. We received 12 automatic acknowledgment, sometimes accompanied with a ticket for further communication or a link to an online portal to provide additional information. Only a little part of these 12 operators then contacted us. Moreover, we received 7 manual replies, one of which reporting a false positive and pointing out a bug in CAIDA's system (which has now been fixed). This operator has then been excluded from the study, as the network was indeed deploying anti-spoofing filters. In the remaining 6 replies, operators told us that they were going to investigate and fix the problem.

Table 5. Notification results.

Group	Country	Spoof ASes	Notification sent	Replies		Website views
				Auto	Manual	
Control	CAN*	19	-	-	-	-
	FRA**	5	-	-	-	-
	TUR***	7	-	-	-	-
Total control group		31	-	-	-	-
Private	AUS*	15	14	1	-	3
	DEU**	11	11	2	2	6
	POL***	4	4	-	1	1
Total private group		30	29	3	3	10
Public	GBR*	22	21	8	3	10
	NLD**	9	9	-	1	3
	ITA***	7	7	1	0	4
Total public group		38	37	9	4	17

Note: * triplet 1; ** triplet 2; *** triplet 3.

5.1.2 Analysis of operators' reaction

We took these 6 replies as a chance to further investigate the reasons why these operators lack anti-spoofing filters, replying to their mail with brief questions.

The majority of operators told us that they are deploying anti-spoofing, and suggested that the test may have spotted some misconfigured interfaces. One operator wrote: *"We have ingress anti-spoofing filters on our customer interfaces and also on our uplinks. I am very surprised that there are IP addresses that can spoof".*
Another added: *"We have been using anti-spoofing mechanism for a long time. This is a very important issue for us. For some reason, some interfaces are not enabled. Verification and repair is in progress".*

There are also few cases in which it seems that networks do not deploy anti-spoofing due to the limitation of uRPF. As you recall from 2.2.5, uRPF (BCP84) is the automatic way to implement Access Control Lists prescribed by BCP38, which overcomes problems with multihomed networks and

simplifies the deployment of filters on large and very dynamic networks. From our previous interview it has emerged that, though uRPF reduces the costs and the efforts to implement anti-spoofing, it may also expose the network to significant drops in performances and old equipment might not supported its implementation.
One operator wrote us: *"We have a solution to this* [talking about the lack of filters] *that will be rolled out over the next few months, as we upgrade our core routers".*
Another operator further explained his situation: *"The network you have mentioned is from our virtual server product. … We are unable to use uRPF in all segments because of hardware compatibility issues and stability. We have now implemented ACL based filtering in all relevant segments as temporary fix. During this year we have planned to upgrade to new hardware which supports uRPF correctly".*

In some cases, operators mentioned to have had previous discussion on the deployment of filters with CAIDA, or mentioned to be member of MANRS, yet pointing out problems in correctly deploying anti-spoofing filters.

We also designed a short questionnaire to investigate operators' feedback. Unfortunately, we got only 2 responses. In one case, the operator reported that anti-spoofing filters were deployed, but loosely configured on some interfaces. Both operators mentioned they were going to fix the problem. In addition, both the operators found our website *somewhat useful*.

As for the visits to our website, it is noteworthy that the link included in the notification has been opened in 27 cases on the 66 notifications sent (40.9%), suggesting that the nudging tone of our notifications might have attracted operators' attention, at least in an initial moment.

5.1.3 Third party engagement

For countries in the public-public treatment, we further contacted the national CERT, the Network Operator Group (NOG) and security blogs.
The Dutch CERT, part of the Dutch National Cyber Security Center (NCSC), immediately replied us asking further details before taking actions and contacting non-compliabt operators. As for the British CERT, our first notification was sent to an old address. After the first week without replies, we contacted the British NCSC via an online form on their website. Few days after we received a reply saying the information has been passed to the relevant team. Finally, we have not received any reply from the Italian CERT.
However, all three these CERTs opened the link to our website in multiple points in time.

Secondly, we tried to engage the community of network operators via the NOGs, achieving a satisfying outcome. We used the mailing list of the UK-NOF (Network Operators Forum), the NL-NOG, and the IT-NOG. As shown in Table 6, NOGs have boosted the visits to the website: in the 25 days of the experiment, 202 unique IP addresses opened our webpage from the UK-NOF, 178 from the NL-NOG and 104 from the IT-NOG.
In addition, we received some emails from operators interested in our website, ranging from positive comments to suggestions to improve the coverage of the measurements. More importantly, during an

email exchange with a German operator, we found out that somebody forwarded our notification to the German NOG. On the one hand, this clearly interferes with the structure of the treatments (at least in Germany). However, on the other hand, it demonstrates that our notification successfully engaged the community of operators and confirmed that NOGs are a valuable tool to disseminate security information among operators. Furthermore, this operator told us that the abuse email address to which we sent the private notification is automatically handled by their system, suggesting another address for this type of security issues. This is an interesting result, as it shows that the abuse field of the WHOIS protocols might have a limited reliability. It may also contribute to explain why many operators did not reacted to our notifications.

Finally, we tried to contact security blogs in the countries in the public-public group in order to further promote our website. Only Secutiy.nl, a Dutch blog, reacted to our email and published a short piece about our website (possibly due to the good reputation of TU Delft in the Netherlands).

Table 6. NOGs visit to our website grouped by page visited.

NOG	Unique IP visiting
UK-NOF	202
NL-NOG	178
IT-NOG	104

Table 7. Visits to our website grouped by page visited.

Website page	Unique IP visiting
Homepage	417
Remediation	176
Australia	95
Germany	375
Italy	169
Netherlands	560
Poland	52
United Kingdom	290

Table 8. Number of unique IP addresses that visited our website grouped by IP country of origin.

Country	Unique IP	Country	Unique IP	Country	Unique IP	Country	Unique IP
AT	7	DK	4	IT	106	SC	2
AU	43	EE	2	KR	1	SE	3
BD	1	ES	2	LU	3	SK	3
BE	13	FR	24	MV	1	UA	2
CA	12	GB	187	NL	504	US	209
CH	11	GI	1	NZ	2	UZ	1
CL	1	GR	1	PH	1	VN	1
CN	13	IE	19	PL	15	ZA	1
CZ	8	IM	3	PR	1		
DE	305	IN	5	RU	14		

5.1.4 Increase in spoofable networks observed

A conclusive detail which is interesting to report is that, during the period of the experiment, we saw an increase in the number of spoofable networks identified by CAIDA's Spoofer Project. Table 9 reports the ASN of the spoofable networks observed after the beginning of our experiment (in parenthesis the number of ASes that appeared also in the visits to our website. It is significant to notice that the majority of these ASes is from Italy and the Netherlands, which were in the public group, suggesting that the involvement of the network operator community had a positive effect on the number of test. In addition, we registered 82 clicks on the link in our website to CAIDA's page for downloading the Spoofer application.

Table 9. Increase in spoofable ASes after our experiment.

Canada	France	Turkey	Australia	Germany	Poland	United Kingdom	Netherlands	Italy
-	-	-	1 (1)	-	-	1 (0)	2 (2)	3 (3)

5.2 Post-test measurements and remediation

After 25 days from the beginning of the experiment, we started evaluating remediation by launching a second round of crowdsourcing measurements and by analysing new tests registered in CAIDA's system.

As you recall from Table 4, our pre-test measurements identified 12 spoofable prefixes.
We launched post-test measurements requesting allowing only those 12 users to do the test. In order to encourage these users, we raised the reward to £2.
Only 8 users took part in the post-test measurements. In 6 cases, these tests confirmed that the prefix was still spoofable after our intervention. In one case, the prefix changed status from spoofable to blocked, revealing that anti-spoofing has been implemented (in Australia). In the last case, the users came a different IP prefix from the one previously tested, so that we were not able to assess remediation of the original prefix.

Next, we checked the status of every AS included in the experiment on CAIDA's website, to further investigate remediation. In total, we found out that prefixes from 5 ASes changed status, presenting evidences of remediation.

No evidence of remediation was found in the control group.

Finally, we considered as remediated those ASes whose operators contacted us, as explained in 4.7. Most of the human replies messages we received included a line like *"Verification and repair is in progress",* suggesting the intention to remediate. Some of these operators have also used the Spoofer application to prove that they indeed deployed filtering. Other operators, instead, did not performed the test, even though we asked. For example, an operator wrote us: *"Thanks for the heads up, we have immediately applied access-lists on the routers causing this issue",* but no tests confirmed it. As argued in 4.7 we decide to trust these operators, and we will consider these ASes as remediated.

Table 10 provides an overview of the remediation between public and private groups (note that since we did not find any evidence of remediation in the control group, we have omitted it from this table). In particular, from the fourth to the seventh column, we show the number of operators that visited our website (column *visited*), the number of operators that contacted us showing intention to remediate (column *intention*), and the number of instances in which remediation has been verified with the Spoofer application (column *tested*). As mentioned above, some operators have contacted us but did not perform the test, while others have remediated without writing us. Therefore, the column *remediated* shows the number of unique ASes that have we consider remediated (which is different from the sum of the two previous columns).

Table 10. Remediation.

Treat	Country	AS notified	Visited	Remediation			Remediation rate
				Intention	Tested	Remediated	
Private disclosure	AUS*	15	3	-	1	1	1/15 (6.67%)
	DEU**	11	6	2	1	2	2/11 (18.2%)
	POL***	4	1	1	-	1	1/4 (25%)
Total private disclosure		30	10	3	2	4	4/30 (13.3%)
Public disclosure	GBR*	21	10	2	-	2	2/21 (9.52%)
	NLD**	9	3	2	2	2	2/9 (22,2 %)
	ITA***	7	4	1	1	2	2/7 (28,6%)
Total public disclosure		37	17	5	3	6	6/37 (16.2%)
Overall total		67	27	8	5	10	10/67 (14.9%)

Note: * triplet 1; ** triplet 2; *** triplet 3.

5.3 Data analysis

5.3.1 Survival analysis

To establish the effect of the of public and private information disclosure on deployment of anti-spoofing filters, we compare the probability of remediation among the treatment groups. To this aim, we use *survival analysis*, a type of statistical analysis that looks at the time before a determinate event happens in order study the portion of the population that "survives" as a function of time.

Methodology

The main focus of survival analysis is on the *survival function*, which expresses the probability of occurrence of the event under investigation at any given time. Survival analysis finds application in many fields. In medical research, for example, the survival function describes patient's mortality in response of different treatments. In engineering, the survival function is known as reliability function, and it is used to assess the reliability of system before a failure happens. In our case, the event observed is remediation. Thus, we going to analyse if the probability that an AS get remediated differs significantly among the different experimental groups. To be more precise, we will compare the survival curves associated with the three treatments, which represent the probability that an AS survives (i.e. is not remediate) over time. In particular, we seek to understand the different effect of private and public disclosure. This difference is not only related to the number of ASes that remediated (i.e. the final outcome), but must also consider differences in the time in which remediation happens (i.e. the speed of remediation for each treatment). Typical regression methods cannot account for these factors,

and might not be able to reveal differences among treatments. Therefore, survival analysis provides the perfect tool for our needs.

To visualise the survival function of a group of observations, or to compare survival functions of different groups, *Kaplan-Meier plots* are used. Kaplan-Meier is a non-parametric methodology used to estimate and plot the survival probabilities over time.

The statistical test used to compare differences in the distribution of probabilities of two (or more) survival curve is called *log rank test*. It can be associated with a large sample χ^2 statistics, where categories are the set of ordered failure times. The idea behind the log rank test is that at each occurrence of the event, an expected occurrence count is computed for each group, given the survival probability in that group at the time the event occurs (Kleinbaum, 1998). The log rank test between two distribution of probability A and B can be formulated as:

$$Log\ rank\ test = \chi^2{}_{log\,rank\ test} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B},$$

where O_A and O_B are the total number of observed events in groups A and B, and E_A and E_B are the expected value of the number of observed events in group A and B.
As for the classical χ^2, we can use the critical value from the χ^2 table. For comparison between two groups (*1 df*) and a significance of 0.05, the critical value is $\chi^2_{0.05,1} = 3.84$. For results of the log rank test over this threshold, the null hypothesis (i.e. no difference between survival curves of two groups) must be rejected.

Results

We begin our analysis by considering the overall effect of our interventions, regardless of the original cluster to which countries belong. As a first step, we plot the survival probabilities of all the ASes included in our experiment, grouped by treatment.

Figure 14 juxtaposes the survival curves of ASes from all the countries in the control group (red line), in the private group (blue line), and in the public group (green line). Firstly, we notice that the survival curves of both treatments appear different from the one of the control group. However, it seems that the difference between the two treatments is minimal.
As mentioned in the previous section, the we did not find any evidence of remediation in the control group, and thus its survival function is a straight line representing a constant survival probability of 100%. As for the private group, after 13 days form the notification 4 operators remediated, bringing the survival probability to 86.7%. As no additional operators remediated, the survival curve after day 13 stayed constant.
A similar trend is also evidenced by the survival function for the public group. In this case, four operators remediated in the first 11 days, after which the probability of surviving was constant at 83.8%.

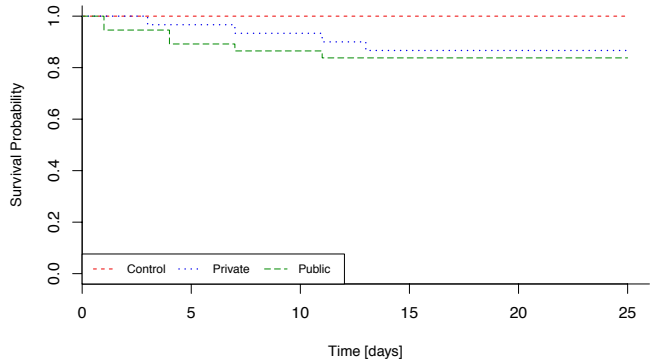


Figure 14. Survival probability for all ASes (no cluster distinction).

Table 11. Log rank test results (no cluster distinction)

Group	Control		Private		Public	
	χ^2	p	χ^2	p	χ^2	p
Control			4.4	0.0369**	5.4	0.0201**
Private	4.4	0.0369**			0.1	0.703
Public	5.4	0.0201**	0.1	0.703		

Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$

To further investigate the difference among survival curves we perform a series of log rang tests between the survival curves of control and private group, control and public group, and private and public group. Table 11 shows the results of the tests, which confirm that there is a small, yet significant difference between the survival curves of control group and private group ($\chi^2 = 4.4$, $p = 0.0369$), and between control and public ($\chi^2 = 5.4$, $p = 0.0201$). However, the test revealed that there is no significant difference between the two treatments ($\chi^2 = 0.1$, $p = 0.703$).

Next, we apply the same typology of analysis (plotting the Kaplan-Meier plots, and testing every combination of treatments with a log rank test) to check if the effectiveness of the treatments is affected by the variables that characterise the clusters of countries (i.e. GDP per capita, ICT Development Index, Global Cybersecurity Index, and whether a country is native English speaker).
Figure 15 shows the survival curves of the three treatments in the first cluster of countries (Canada, Australia and the United Kingdom), Figure 16 refers to the second cluster (France, Germany and the Netherlands), and Figure 17 to the third (Turkey, Poland and Italy).
For each cluster, we also present the results of the log rank tests respectively in Table 12, 13 and 14, which show that no significant difference is found among the treatments when grouped by country clusters. Moreover, in Figure 18, the survival curves of each cluster are juxtaposed, and Table 15 further shows that remediation does not vary significantly among clusters.

Figure 15. Survival probability (first cluster).

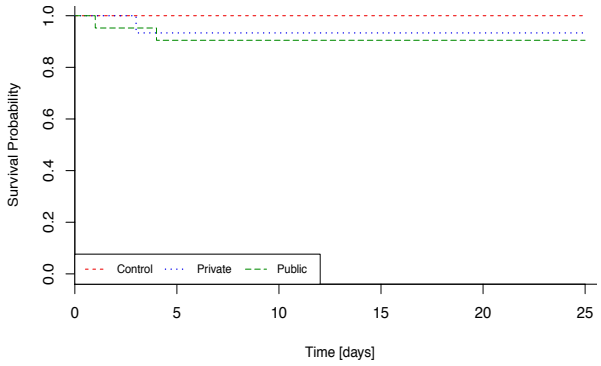


Figure 16. Survival probability (second cluster).

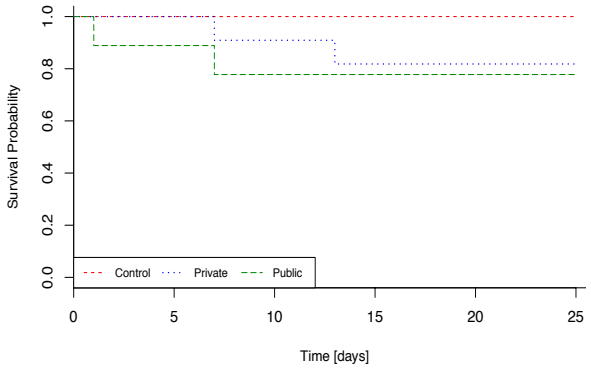


Figure 17. Survival probability (third cluster).

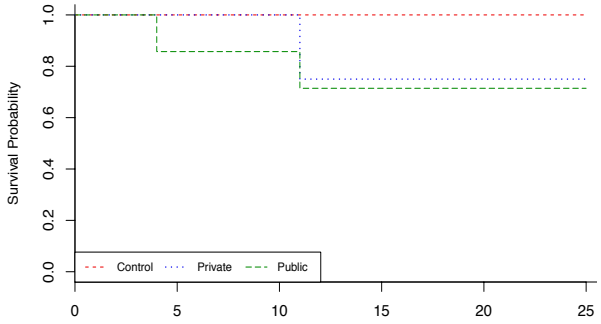


Figure 18. Survival probability grouped by cluster.

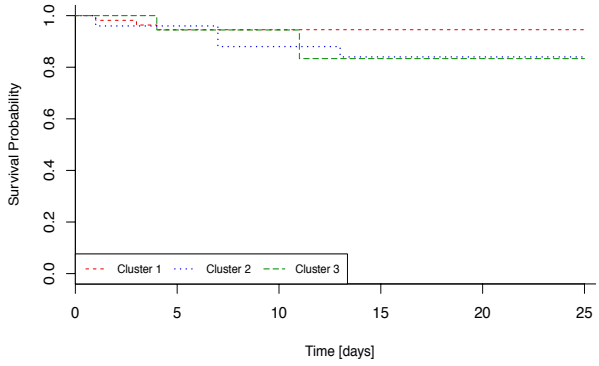


Table 12. Log rank test results for the first cluster.

Group	Control		Private		Public	
	χ^2	p	χ^2	p	χ^2	p
Control			1.3	0.26	1.9	0.173
Private	1.3	0.26			0.1	0.764
Public	1.9	0.173	0.1	0.764		

Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$

Table 13. Log rank test results for the second cluster.

Group	Control		Private		Public	
	χ^2	p	χ^2	p	χ^2	p
Control			1	0.329	1.2	0.277
Private	1	0.329			0.1	0.761
Public	1.2	0.277	0.1	0.761		

Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$

Table 14. Log rank test results for the third cluster.

Group	Control		Private		Public	
	χ^2	p	χ^2	p	χ^2	p
Control			1.8	0.186	2.2	0.141
Private	1.8	0.186			0	0.84
Public	2.2	0.141	0	0.84		

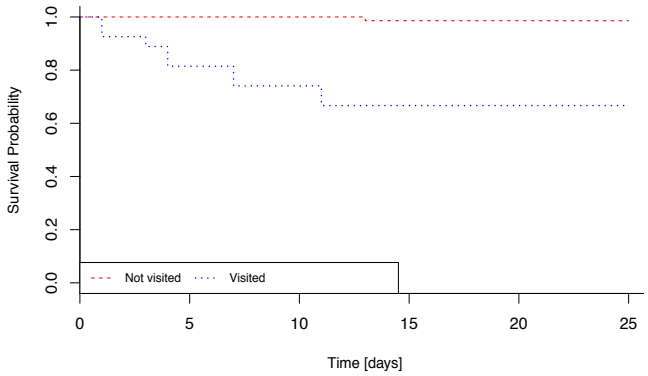
Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$

Table 15. Log rank test results (grouped by cluster).

Group	Cluster 1		Cluster 2		Cluster 3	
	χ^2	p	χ^2	p	χ^2	p
Cluster 1			2.2	0.134	2	0.154
Cluster 2	2.2	0.134			0	0.971
Cluster 3	2	0.154	0	0.971		

Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$

Figure 19. Survival probability for ASes who visited the website.



We conclude the first part of our analysis by examining the effect of viewing our website on the chance of remediation. To this aim, Figure 19 shows the survival function for ASes that opened the link to our website and who did not. It appears evident that operators that opened our website are more likely to remediate. This observation is also corroborated by the results of the log rank test ($\chi^2 = 23.7$, $p = 1.12e\text{-}06$).

5.3.2 Regression analysis

After analysing the effectiveness of the notification with survival analysis, we now seek to provide an additional insight on two findings of the experiment: the number of operators that visited our website and the moderate remediation rate. In particular, we seek to identify some characteristics of the ASes that might predict the likelihood that notified operators open the link to our website and the likelihood of remediation. To this aim, we perform two regression analysis. However, since we failed to observe significant effects of the predictors on the visit to the website, in what follows we are only discussing the analysis of the occurrence of remediation. The analysis of the website visits is reported in Appendix 10.

Methodology

Since both the visit to our website and remediation, the two dependent variable, are binary (i.e. 1 if the AS remediated, 0 otherwise), we apply logistic regression to model them.
An advantage of logistic regression is that both continuous and discrete variable can be used, as there is no restriction on type of predictors. Moreover, logistic regression analysis does not require that the predictors have a normal distribution (Cetin et al., 2017).

We use the following variables to predict the occurrence of remediation:

- *Organisational factors*
  - ***x_1: AS size***
    A continuous variable measured via the number of IP prefixes announced by the AS.
  - AS type
    A categorical variable for the type of the AS, divided in the following binary variables:
    - ***x_2: Access provider***
    - ***x_3: Enterprise***

    - ***x_4: Content provider***
- *Socio-technical factors*
    - ***x_5: GDP per capita***
      The GDP per capita of the country of the AS.
    - ***x_6: ICT score***
      The score of the country of the AS on the ICT Development Index.
    - ***x_7: GCI score***
      The score of the country of the AS on the Global Cybersecurity Index.
    - ***x_8: English native country***
      A binary variable set to 1 for ASes in English native speaker countries.
- ***x_9: Visited***
  A binary variable set to 1 for operators who visited the website.

The equation of the logit model is:

$$logit(\pi_r) = ln\left[\frac{\pi_r}{1-\pi_r}\right],$$

where π_r represents the occurrence probability of remediation, modelled as:

$$\pi_b = \frac{exp(\beta_0 + \sum_i \beta_i x_i)}{1 + exp(\beta_0 + \sum_i \beta_i x_i)},$$

where x_i, $(i = 1, \ldots, 9)$ refers to the nine variables under investigation, β_i is the partial regression coefficient, and β_0 is the intercept. The various terms $exp(\beta)$, are *odd ratios* that reflects the correlation between visits to the website and the probability of remediation. Values of $exp(\beta) < 1$ suggest a negative correlation ($\beta < 0$). If $exp(\beta) = 1$, (i.e. $\beta = 0$), the two variables are not correlated. For value of $exp(\beta) > 1$, a positive correlation exists ($\beta > 0$).

The sample on which we perform the analysis includes the 67 ASes notified, both privately and publicly, excluding those in the control group that did not received any notification.

Appendix 11 we report the R code for the logistic analysis

Modelling occurrence of remediation

The results of the logistic regression are presented in Table 16. Only the variable *visited* has a significant effect on the remediation rate. The coefficient for the variable visits is $\beta_9 = 2.32$, and its odds ratio is $exp\ (\beta_9) = 10.23$, meaning that the odds of remediation for operators who open the website are 10.23 times larger than the odds of remediation of operator who does not (odd ratio: 10.23, confidence interval: [0.37, 4.28]).

To assess the goodness-of-fit of the model, we plot the Receiver Operating Characteristic Curve (ROC), shown in Figure 20. It summarizes the model performance between sensitivity (true positive error rate) and specificity (false positive error rate). Next, we compute the Area Under the Curve, which reveals that the model is very accurate (with 96% AUC score).

Finally, we compute different pseudo-R^2 parameters, shown in Table 17. McFadden and Negelkerke pseudo-R^2 are fairly high, confirming the good fit of our model.

Table 16. Results of logistic regression analysis.

	Dependent variable	
	Remediation	
x_1: AS size	0.25	(1.08)
x_2: Access provider	7.61	(1923.09)
x_3: Enterprise	6.87	(1604.86)
x_4: Content provider	7.60	(1371.92)
x_5: GDP per capita	0.06	(1.23)
x_6: ICT score	-1.74	(1.45)
x_7: GCI score	3.11	(2.08)
x_8: English native country	-2.95*	(1.76)
x_9: Visited	2.32**	(0.99)
Observations	67	
Log-likelihood	-11.41977	

Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$; Standard error in brackets

Figure 19. Model diagnosis with ROC curve.

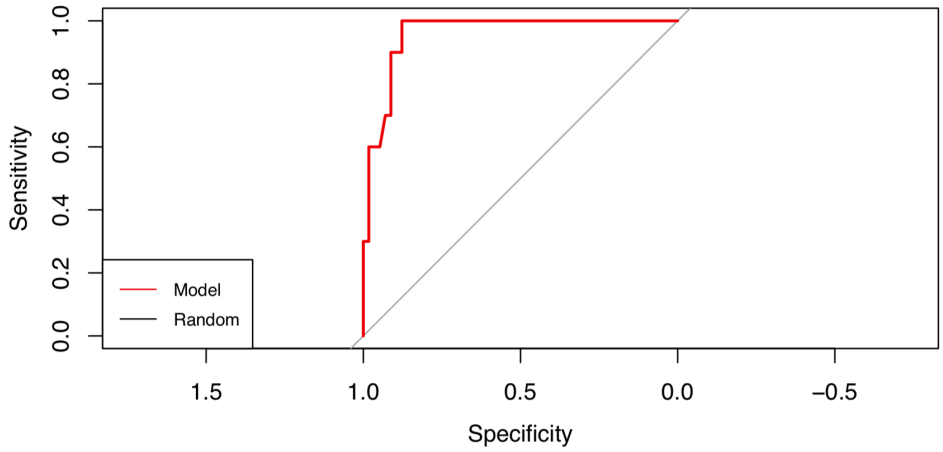


Table 17. Goodness-of-fit.

	Pseudo-R²		
AIC	McFadden	Cox & Snell	Nagelkerke
42.8395375	0.595	0.395	0.693

5.4 Discussion

In the previous sections, we presented and analysed the results. But how does this analysis contribute to answer our research questions? In this section, we discuss our results in relation to the hypothesis formulated in Section 3.4. Next, we compare our results with the findings of previous studies.

5.4.1 Hypothesis validation

Hypothesis 1:
Operators privately notified are more likely to remediate than operators not notified.

The survival probability of ASes in the in the private group is significantly different from the one of ASes in the control group (Figure 14, Table 11). Therefore, we can accept our hypothesis of a positive effect of private notification on the likelihood of remediation.

Hypothesis 2:
Operators publicly notified are more likely to remediate than operators not notified.

Similarly to the previous hypothesis, the analysis of the survival probability in the public and control group revealed that the remediation rate of publicly notified ASes is significantly different from the one of ASes not notified (Figure 14, Table 11). Thus, we conclude that public notifications have a positive effect on the likelihood of remediation, accepting our hypothesis.

Hypothesis 3:
The likelihood of remediation varies significantly between operators privately and publicly notified.

The survival analysis has not showed any significant difference in the probability of remediation of ASes in the private group compared to those in the public group. Thus, we must reject our hypothesis, and accept the null hypothesis that the two groups have the same survival curve.

Hypothesis 4:
Organisational and socio-technical factors affect significantly the effectiveness of notifications.

Comparing the survival probabilities of ASes in different country cluster has not shown any significant difference in the remediation rate. Moreover, according to the regression analysis, organisational and socio-technical factors have not a significant effect on the likelihood of remediation. Thus, we conclude that neither the organisational nor the socio-technical factors we selected affect the effectiveness of notification, rejecting this hypothesis.

Hypothesis 5:
Operators who visit our website are more likely to remediate than those who do not.

The survival curve showed that the remediation rate of operators who visited our website at the end of our study was significantly higher than those who did not.

In addition, the logistic regression identified the visit to the website as a good predictor of remediation. Therefore, we accept this hypothesis.

5.4.2 Discussion of the results

In this section we review our results, and we discuss them in relation to the findings of previous notification experiments.

On the 67 AS notified, 10 have remediated (14.9%). In particular, 4 ASes on 30 (13.3%) remediated in the private group, and 6 on 37 (16.2%) in the public group, while no evidence of remediation has been observed in the control group. Though modest, these remediation rates show that our notifications had an impact on the deployment of filters.

We did not find any significant difference in the remediation rate of ASes publicly and privately notified[20]. This result can be explained in two ways.
First, it might be that reputation and peer pressure do not represent strong incentives for operators to deploy filter. Secondly, it might be that our experiment did not manage to create a "reputation effect" big enough to engage network operators.
We are more incline to think that the lack of difference between public and private notification is due to this second explanation. In fact, previous research found a significant effect of peer pressure and reputation effect in the reduction of spam among organisation in the same industry (He et al., 2016). In particular, the public notification of Tang et al. (2013) induced a 15.9 reduction of spam, which is consistent with our findings.

We believe that the time span of experiment and the intensity of the promotion of the initiative are key elements that can explain our limited results.
We think that more time is needed to set in motion the effect of reputation and peer pressure or, at least, to observe the consequence of such effect.
Similarly, the effectiveness of public notification might depend on the visibility of the information disclosed. A more targeted approach the promotion of our website (e.g. via social media), together with additional reminders to non-compliant operators could be beneficial in this sense.
Finally, the effectiveness of public disclosure depends also on the relevance of the information disclosed. In our case, despite information about networks lacking adequate anti-spoofing measures might be important in a regulatory context, it has not a great relevance for defenders (e.g. does not enable other operators to improve their security).

In terms of results, our 14.9 remediation rate is in line with the findings of previous research. Researchers of CAIDA found that private notification to non-compliant operators induce 15-20% of operators to remediate.
When we contacted CAIDA at the end of our experiment, we discovered that the 95 on the 98 of ASes we notified already received a private notification. Around 15% of the ASes included in our experiment

[20] Despite the survival probability in the two groups are similar, we noticed that remediation appears to occur quicker in the public group. However, this difference might also be due to chance.

received a notification in 2018. Of these, only few received a notification during our experiment. Interestingly, 2 of the ASes notified by CAIDA during our experiment remediated. Moreover, before and during our experiment, researchers of CAIDA also sent notification to operators in our control group, but no remediation has been observed there.

Operators' feedback to our notifications has been minimal.
We collected 6 manual replies, and a questionnaire included in the notification only got two responses. The majority of operators seemed unaware about the problem. In particular, three operators told us they were already deploying anti-spoofing, and reported minor errors in the configurations of the filters. Moreover, two operators reported technical limitations: one mentioned problem with fragility of the network caused by uRPF, and another mentioned the need to upgrade the core routers. It is interesting to notice that these answers reflect also the information collected with our preliminary interviews. All in all, it seems that the challenge of IP spoofing is still affected by technical limitations.

Finally, two important results of our experiments refer to the engagement of the network operators with our website.

First, we found that 40.9% of the operators notified opened the link to our website. Previous notification studies which included a demonstrative website recorded a very little engagement of participants. When investigating the role of sender reputation on the effectiveness of notifications, Cetin et al. (2016) found that only 8% of participants opened a demonstrative website. Likewise, in a second study surveying different notification strategy, Cetin et al. (2017) observed that less than 15% of participants opened the website included in the notification, suggesting that the bottleneck is in getting recipient of the notification to visit the website.
We attribute the success of our website to the nudging tone of our notification. Moreover, we also showed that the visit to our website is a good predictor of remediation. Thus, we conclude that providing social information about the behaviour of people in a community is a good way to attract their attention to the problem. As mentioned in Chapter 3, this is an important finding of behavioural economics, and we argue that can be further applied to security notifications.

Secondly, our notification to the NOGs managed to engage the network operator community. Besides boosting the number of visits to our website, some operators supported us by sharing our webpage with other closed industry groups, showing that the community is sympathetic to the problem and available to help. Therefore, we argue that researchers should make use of this collaborative environment.
The fact that cooperation and information sharing is beneficial for Internet security is not a new idea. For example, Kührer and colleagues (2014b) sent notifications about a vulnerability exploited during NTP amplification DDoS, and achieved 92% remediation by sharing vulnerability information with different actors in the security community (CERTs, network operators, hardware manufactures, data clearing houses).

As a final note, one month after the end of the experiment we checked again the presence of new Spoofer test on the 98 ASes included in our study to evaluate the consistency of our initial analysis.

We noticed that for 55 ASes no new tests have been collected after the begin of our experiment on the IP prefixes we reported. However, for 4 ASes new tests were performed on IP prefixes close to those we reported (/23 and /22), showing the presence of anti-spoofing filters. We believe it is possible that these 4 ASes remediated. Interestingly, two of these ASes are in the United Kingdom, one is in the Netherlands and one is in Italy, which are all countries that were in our public group.

In 27 cases, new tests revealed the lack of filters, confirming that the AS was still spoofable. One of these tests came from an AS we have considered remediated (as the operator told us that the problem was being investigated).

Finally, 3 ASes presented evidence of remediation after the end of our experiment (one in the Netherldans, one in Australia and one in Canada).

5.5 Conclusions

In this chapter we have presented and analysed the results of our experiment. Of the 66 notified operators, 27 opened our website. Moreover, we received 12 automatic acknowledgements and 7 manual replies (some from ASes from which we also received an automatic acknowledgement). For countries in the public group, information about non-compliant operators has been shared with the national CERTs, the NOGs and security blogs. Two CERTs on three replied us. The promotion of our website via NOGs was fairly successful, and a high number of operators opened the link we provided. Finally, only one security blog accepted the request of posting about our website, which led to an additional increase in the number of visit to our website.
Next, we analysed the remediation rate. We observed remediation in 10 cases on 67 (14.9%). In particular, 4 ASes on 30 (13.3%) remediated in the private group, and 6 on 37 (16.2%) in the public group. No evidence of remediation has been observed in the control group.
Lastly, the survival analysis showed a significant effect of notification on the remediation rate, though no difference between public and private notification has been observed. Moreover, we did not find evidences of the effect of organisational and socio-technical factors on the likelihood of remediation. However, the regression analysis showed that operators who visited our website are more likely to remediate.

Chapter 6:
Discussion and conclusions

6.1 Answering the research questions

6.1.1 Reviewing the research questions

The problem of incentivising operators to deploy anti-spoofing filters lied at the heart of this research. As previous attempts of privately reporting the lack of filters to operators found a moderate remediation rate, we sought to find alternative approaches to notify non-compliant operators, and engage them in remediating the problem. To this aim, in this research we have proposed the use of private and public disclosure of security information: we designed a website to aggregate measurements of compliance with anti-spoofing, and to display which operators are found without proper defences. We used this website to notify a group non-compliant operator, in order to test the effect of privately disclose compliance information on the adoption of anti-spoofing filters. Additionally, a second group of operators was exposed to the public disclosure of the same information, and relevant third parties had been notified to create additional incentives out of reputation and peer pressure. We were also interested in understanding the role of the organisational and socio-technical factors in explaining differences in the likelihood of remediation. Finally, we wanted to investigate the factors that prevent operators from correctly deploy anti-spoofing filters, in order to formulate recommendation for designing future interventions tailored to operators' perception of the problem.

All these elements contribute to answer the main question of this research:

- *To what extent do notifications incentivise compliance with anti-spoofing best practices?*

In the following sections, we review our results and we formulate the answers to the sub-questions we need to address, in order to then answer our main research question.

6.1.2 Private notification

1. *What is the effect of privately notifying non-compliant operators?*

By comparing the overall remediation rate of operators exposed to the private disclosure with the remediation in the control group, we see that our notifications provoked a remediation in 4 ASes on 30 (13.3%) (Table 10).
Two operators remediated after contacting us (one from Poland and one from Germany). One case of remediation was only measured via the Spoofer test, without any communication from the operator. In the final case, the operator contacted us, not because of our initial notification, but because our website has been posted on the German NOG without our knowledge.

This shows an inherent limitation of private notifications: finding reliable contact points. In our study, we used the abuse contact retrieved via WHOIS lookup. We observed that 10 operators on 30 opened our website. Why 20 operators did not visit the website remains unknown. It might be because the address to which we directed our notification was incorrect, or because operators have ignored the message. At the same time, if 10 operators visited the website, we also wonder why only 3 contacted us (at least one of which via the NOG). Again, this might be because the addressee was incorrect. Another speculation is that our information was unreliable. In fact, for some operators the last spoofable test was fairly old, and they might have since remediated (however this does not explain why they did not contacted us to correct the information we display on our website). These evidences seem supporting the idea that operators do not care about IP spoofing, and that most of operators have ignored our mail notification.

In spite of that, the results of the survival analysis showed that the remediation rate in the private group is significantly different from the remediation in the control group. Therefore, we conclude that private notifications have a moderate, yet positive impact on the deployment of anti-spoofing filters. We argue that the small impact of private notification is due to the difficulties of finding reliable contact information of the party to notify.

6.1.3 Public notification

2. *What is the effect of publicly notifying non-compliant operators?*

For countries assigned to the public disclosure group, remediation was marginally better: 6 operators on 37 remediated (16.2%). Also the visits to our website (17/37, 45.9%) were slightly higher than in the private group. Still only 5 operators contacted us (plus one case in which remediation was observed looking at CAIDA's test, without any communication from the operator). However, we still remain with the same unanswered questions posed above, namely: why 54% of the operators did not open our website, and why only 5 operators contacted us?

As in the case of the private notification, the remediation rate in the public group differs significantly from the control group. However, the remediation rate and all the other indicators shown no significant difference in remediation between private and public group. This makes us speculate that the difference between private and public disclosure is marginal. At least in terms of remediation rate.

In fact, publicly disseminating vulnerability information might attract the attention of the vulnerable party.
This is exemplified by the case of the German operators who contacted us via the DE-NOG, who wrote us *"Could you please change out status to remediating on https://www.infospoofing.com/de? Would be great to be not first place in this list!"*. Interestingly, this reply shows the reputation effect we wanted to produce on an operator that did not received the private notification. Thus, we argue that using public mailing list, like the NOGs, is a valuable solution for reporting abuse. In addition, we received different mail from operators and researchers interested in our initiative, who further helped us in promoting our website. This result suggests that public notification, if well crafted, had the potential of engaging the community, and this engagement can open the doors to new notification strategies.

6.1.4 Role of operators' characteristics

3. *Can we identify characteristics of network operators that explain differences in remediation?*

Our experiment was designed to observe if the socio-technical characteristics of a country affect the effectiveness of the notifications. By comparing the remediation rate of operators in different country clusters, we have not found any such effect. Moreover, we have also analysed whether organisational factors such as the type and the size of an AS affect the likelihood of remediation. Also in this case, we have not found evidences supporting this relation. Therefore, we conclude that organisational and socio-technical factors have no effect on the likelihood of remediation.

6.1.5 Recommendations

4. *What practical recommendations can be formulated on the base of the previous findings?*

On the base of main observations emerged during our research and given the results of our experiment, we identify seven practical recommendations, five addressed to security researchers and two addressed to network operators.

We begin with the recommendations to security researchers.
The first remark is that many operators did not reacted to our website, not even by clicking the link. This fact makes us wonder about the reliability of the contact information retrieved via the WHOIS protocol. As the supply of vulnerability is theoretically endless, so vulnerability notification will accompany cybersecurity researchers for a long time. In light of this, a fundamental obstacle to effectiveness notification is finding the right contact information. Our first recommendation therefore is:

- *To improve WHOIS contact reliability, for example by reporting incorrect addresses in order to require updates of the abuse contacts.*

Improving the WHOIS reliability might represent a big step forward for the notification process. Though alternative methods to facilitate information sharing exists, they are likely to be based on voluntary participation, and free ride might occur (i.e. participants tent to receive information without sharing). In its way, the problem with WHOIS contacts also undergoes a similar type of externality: a party may rely on WHOIS for retrieving contacts about other, not about himself, so it might be that this party has no incentives to provide accurate information about himself. WHOIS represents our canary in the coal mine. If we cannot fix it, it will be difficult to achieve the degree of cooperation at the base of cybersecurity.

Alternatively, another approach might be to gradually abandon direct mail notification, and focus instead of other type of media. In this regard, public disclosure appears suitable. However, reputation effect might require time, as well as conducting an extensive promotion that can move a greater incentive. We attempted to set in motion such mechanism achieving moderate success. Therefore, our second recommendation is:

- *To periodically release information about recent results of the Spoofer test, via NOG and similar public channels, in order to reinforce our initial effort.*

In fact, studying the effect of disclosure policies in the long run may reveal mechanisms that did not manifest in the short time span of our experiment. Note that this recommendation applies to the case of IP spoofing as will as other security issues: publicly disseminating vulnerability information might not only induce remediation via reputation and peer pressure, but it can also increase the chance to attract the attention of the vulnerable party.

Next, the high percentage of operators who opened the website in our notification suggest us that the nudging tone of our notification attracted operators' attention. Given that previous research claimed that getting the recipient of the notification to open a demonstrative website represents a bottleneck for the notification process, we recommend:

- *To craft future notification message keeping in mind insights from behavioural economics and psychology, in order to increase the chance engaging the recipient.*

As for the mitigation of IP spoofing, it is necessary to improve our point of view on the problem from at least two perspectives.
First, we need to better understand the reason why operators do not deploy filters. Our research provided a little insight on this. However, it is important to notice that the operators who contacted us are those that were inclined to remediate, whether because our notifications provided an extra incentive or just because they were unaware about the existence of the problem. In both cases, this small sample cannot be a solid base for generalizations. Arguably, the operators who contacted us are also those more concerned with Internet security or with their reputation. Thus, understanding the reasons why operators appears to be without filters remains an open question for the majority of the operators notified, and we recommend:

- *To engage non-deployer operators, in order to understand the reasons of their lack of compliance and their scares engagement with notification.*

Reaching out to these operators, in a way or another, would clearly contribute to our ability to design tailored strategies to prompt compliance.

Secondly, there is the need of improving our ability to observe compliance with anti-spoofing filters, in terms of network coverage and frequency of measurements. We recommend:

- *To promote the Spoofer application, also outside of the technical community, in order to improve the coverage of the measurements and awareness of the problem.*

Enlarging the set of available measures might allow a better profiling of non-compliant network, and it would lead to more focused interventions. In this regard, using established network measurement

platforms like RIPE Atlas[21] to host the Spoofer test can provide significant improvement to the measurements of compliance.
Moreover, crowdsourcing marketplace appear an interesting approach to recruit volunteers to run the Spoofer test. In our research, we showed that using crowdsourcing platforms to conduct country-specific measurements is a feasible solution, which enabled us to reveal spoofable network not previously identified. Finally, the use of crowdsourcing might also have a positive effect in terms of awareness about IP spoofing among regular Internet users. In fact, some users that took part in our crowdsourcing measurement showed curiosity about the problem (asking for instance how to interpret the results of the test). And since more than 1700 users opened the website we designed to host the crowdsourcing study, we believe that showing these users the results of the measurements on the website we used to notify operators might have been another way to further promote our initiative.

Lastly, we formulate two recommendations for network operators.
The first originate from the analysis of the data collected by the Spoofer Project, and from operators' feedback to our notifications. By analysing the results of the measurements of compliance with anti-spoofing best practices, we noticed that around 70% of the IPv4 test packed were rewritten by a NAT. Though NAT was not designed as a security mechanism, in practice it also prevents some form of spoofed DDoS attack[22]. However, as IPv4 (the current version of the Internet Protocol) is gradually been replaced with IPv6, the use of NAT for preventing IP spoofing appears problematic. In fact, IPv6 does not support NAT. For this reason, we can expect a comeback of the problems related to IP spoofing in the coming future.
Secondly, part of the operators who contacted us reported an error in the configuration of anti-spoofing filters.
For these two reasons, we recommend to network operators:

- *To periodically review the state of their network, in order to prevent their security mechanism from becoming obsolete.*

Finally, given the positive reaction of the community of network operators to our website, we recommend operators:

- *To capitalise on the compliance with security best practices, in order to improve their brand image.*

Showing compliance with anti-spoofing (as well as other security best practice) can be a strong message for customers competitors and other stakeholders about the level of security of an operator. Moreover, it can also contribute to create a culture of security that is clearly beneficial for the Internet ecosystem. In this regard, participating in the MANRS initiative can be a good starting point to improve reputation and trustworthiness.

[21] https://atlas.ripe.net
[22] Note that there are some scenarios in which NAT does not protect from attacks based on IP spoofing.

6.1.6 Private and public disclosure to improve cybersecurity?

To what extent do notifications incentivise compliance with anti-spoofing best practices?

Overall, the effect of the notification on the deployment of anti-spoofing filters was significant, leading 14.9% of the notified operators to remediate. However, we did not find any meaningful evidence of difference between the type of notifications. Indeed, it seems that both private and public notifications had the same effect on the likelihood of remediation, suggesting that it was the notification itself that was effective, rather than the way in which it was delivered.

Moreover, we saw a good engagement with our website: 40% of the notified operators opened the link included to our notifications. This is a satisfying result when compared with the visits to demonstrative website built in previous studies, which was below 15%. We attribute this result to the nudging tone of our notification, designed to grasp operators' attention. However, despite a remarkable result, most of the operators that visited the website did not remediated.

The low effectiveness of private disclosure might be explained in light of the limited reliability of the abuse contacts, or due to lack of care of the notified operators. We argue that both cases are related to a lack of incentives of network operators to adopt RFC best practices.

Though we did not observe any significant effect of reputation and peer pressure on the remediation rate, our public notifications lead to some interesting results. We observed a positive reaction to the disclosure from the operators' community (in terms of visits to our website and engagement), which suggest that, after all, public disclosure has potential, but our intervention was not able to fully reveal it. Remediation aside, public disclosure appears to provide side benefits in terms of increase awareness (for example we saw an increase in the number of spoofable networks during the experiment, probably due to the promotion of our website).

Finally, our results show that the effectiveness of notification is not affected by the characteristics of the notified party. In fact, we have not seen any significant effect of organisational and socio-technical factors on the likelihood of remediation.

All in all, we can conclude that notifying operators has a moderate effect, still positive, on operators' incentives to deploy anti-spoofing filters. Yet, most of the operators ignored our notification, and only a small part was engaged in the remediation of the problem, arguably those more concerned with security. Nevertheless, public disclosure might represent a promising approach for future research to tackle the problem of unreliable contact information.

6.2 Limitation

There are a series of limitation affecting the validity of the results of our experiment.
Field experiment often used to test the effect or proposed policy interventions, since the real setting of the experiment guarantees a high external validity (i.e. generalisability of the results). The flip side of

the coin is that, due to the real setting, internal validity (i.e. confidence in the causal mechanism) might be threatened.

In terms of internal validity, the main limitation affecting our experiment is that we do not know for sure if it was our notification that stimulated remediation. In fact, researchers of CAIDA have sent private notification to spoofable ASes both before and during our experiment. Thus, it might be that operators we counted as remediated had already planned to deploy filters.

As for the external validity, the main limitations refer to the selection of the countries to include in the experiment, and to the lack of rigorous randomisation. Due to the use of crowdsourcing, our selection of countries was constrained by the geographical coverage of the platform. Our list of countries was limited to a group of developed countries in the wester word (with the borderline case of Turkey). A broader selection of country might have led to different results. Secondly, the process of assigning ASes to the experimental conditions was not completely random, and this might affect the external validity of our results.

Next, there are additional limitation of our experimental design. The sample of ASes notified was relatively small, thus affecting the statistical power of our analysis. Moreover, the limited time span of our experiment might not have been enough to observe the effect of reputation and peer pressure. Another factor that might have limited the effect of reputation and peer pressure is the visibility of our website. In this regard, a more constant and intense promotion might have increased the effect of public notifications.

Last, but definitely not least, the availability of data represents a major constrain to our work, which also impact our confidence in the metrics we formulated to determine whether an AS is spoofable and to measure remediation.

6.3 Future research

We conclude by suggesting possible directions for future research.
Despite private and public notification induced a similar remediation rate, our results show that public notification can have some advantages. For this reason, future research should keep investigating whether public disclosure can indeed generate additional incentives out or reputation concern and peer pressure. Our research was among the firsts to take this approach, and our pilot experiment might open the doors to better strategies to notify vulnerable parties. In particular, we suggest experimenting with the use of social media to increase the visibility of the information disclosed. In addition, also suggest future public disclosure intervention to consider a longer time frame, and to accompany the disclosure with periodic reminders and updates to attract attention.

As for the case of IP spoofing, we recommend conducting additional experiments to increase the coverage of the Spoofer data by using crowdsourcing marketplaces. Moreover, we advise researcher in this field to contact operators found without filters to investigate the reasons of their lack of compliance, for example via telephonic interviews.

References

Acquisti, A. (2009). Nudging privacy: The behavioral economics of personal information. *IEEE security & privacy*, *7*(6).

Acquisti, A., Friedman, A., & Telang, R. (2006). Is there a cost to privacy breaches? An event study. *ICIS 2006 Proceedings, 94*.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in Economics* (pp. 235-251).

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The quarterly journal of economics, 94(4), 749-775*.

Allen, V. L., & Wilder, D. A. (1977). Social comparison self-evaluation and conformity to the group.

Asllani, A., White, C. S., & Ettkin, L. (2013). Viewing cybersecurity as a public good: The role of governments, businesses, and individuals. *Journal of Legal, Ethical and Regulatory Issues, 16(1), 7*.

Anderson, R. (1993). Why cryptosystems fail. In *Proceedings of the 1st ACM Conference on Computer and Communications Security* (pp. 215-227). ACM.

Anderson, R. (2001). Why information security is hard-an economic perspective. In *Computer security applications conference, 2001. proceedings 17th annual* (pp. 358-365). IEEE.

Anderson, R. (2010). *Security engineering: a guide to building dependable distributed systems*. John Wiley & Sons.

Anderson, R., Böhme, R., Clayton, R., & Moore, T. (2009). Security economics and European policy. In *Managing information risk and the economics of security* (pp. 55-80). Springer, Boston, MA.

Anderson, R., & Moore, T. (2006). The economics of information security. *Science*, *314*(5799), 610-613.

Anderson, R., & Moore, T. (2007). Information security economics–and beyond. In *Annual International Cryptology Conference* (pp. 68-91). Springer, Berlin, Heidelberg.

Arora, A., R. Telang, & H. Xu. (2004). Timing disclosure of software vulnerability for optimal social welfare. *Proceedings of the 3rd Workshop of Economic Information Systems*, Minneapolis, MN 1–47.

Arora, A., Krishnan, R., Telang, R., & Yang, Y. (2010). An empirical analysis of software vendors' patch release behavior: impact of vulnerability disclosure. *Information Systems Research*, 21(1), 115-132.

Baker, F. & Savola, P., (2004). Ingress Filtering for Multihomed Networks. RFC3704, IETF BCP84.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., & Kraut, R. E. (2004). Using social psychology to motivate contributions to online communities. In Proceedings of the 2004 ACM conference on Computer supported cooperative work (pp. 212-221). ACM.

Beverly, R., & Bauer, S. (2005). The Spoofer project: Inferring the extent of source address filtering on the Internet. In Usenix Sruti (Vol. 5, pp. 53-59).

Beverly, R., & Bauer, S. (2007). Tracefilter: A tool for locating network source address validation filters. USENIX Security (Poster).

Beverly, R., Berger, A., & Hyun, Y. (2009). Understanding the efficacy of deployed internet source address validation filtering. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement (pp. 356-369). ACM.

Beverly, R., Koga, R., & Claffy, K. C. (2013). Initial longitudinal analysis of IP source spoofing capability on the Internet.

Black, J., Hopper, M., & Band, C. (2007). Making a success of principles-based regulation. Law and financial markets review, 1(3), 191-206.

Böhme, R., & Schwartz, G. (2010). Modeling Cyber-Insurance: Towards a Unifying Framework. In WEIS.

Braithwaite, J. (1989). Crime, shame and reintegration. Cambridge University Press.

Buchanan, J. M., & Stubblebine, W. C. (1962). Externality. In Classic papers in natural resource economics (pp. 138-154). Palgrave Macmillan, London.

Butler, B. S. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. Information systems research, 12(4), 346-362. CAIDA, (2018). The Spoofer Project [Website]. Retrieved March 11, 2018 at https://www.caida.org/projects/spoofer/

Chen, Y., Harper, F. M., Konstan, J., & Li, S. X. (2010). Social comparisons and contributions to online communities: A field experiment on movielens. American Economic Review, 100(4), 1358-98.

Chatterji, A. K., & Toffel, M. W. (2010). How firms respond to being rated. Strategic Management Journal, 31(9), 917-945.

Claffy, K.C. (2017). Software Systems for Surveying Spoofing Susceptibility [PowerPoint slides]. Retrieved March 18, 2018 at https://www.caida.org/publications/presentations/2017/software_systems_surveying_spoofing_dhsrd/software_systems_surveying_spoofing_dhsrd.pdf

Camp, L. J., & Wolfram, C. (2000). Pricing security. In Proceedings of the CERT Information Survivability Workshop (pp. 31-39).

Campbell, K., Gordon, L. A., Loeb, M. P., & Zhou, L. (2003). The economic cost of publicly announced information security breaches: empirical evidence from the stock market. Journal of Computer Security, 11(3), 431-448.

Cetin, O., Ganán, C., Korczynski, M., & van Eeten, M. (2017). Make notifications great again: learning how to notify in the age of large-scale vulnerability scanning. In 16th Workshop on the Economics of Information Security (WEIS 2017).

Cetin, O., Hanif Jhaveri, M., Gañán, C., van Eeten, M., & Moore, T. (2016). Understanding the role of sender reputation in abuse reporting and cleanup. Journal of Cybersecurity, 2(1), 83-98.

Collins, R. L. (2000). Among the better ones. In Handbook of social comparison (pp. 159-171). Springer, Boston, MA.

Daigle, L., (2004). WHOIS Protocol Specification. RFC 3912. IETF Standards Track

Damas, J. L. S., (2008). Network Hygiene Pays Off - The Business Case for IP Source Address Verification. [White paper]. Retrieved February 20, 2018 at https://www.ripe.net/publications/docs/ripe-432

Dillenberger, D., & Sadowski, P. (2012). Ashamed to be selfish. Theoretical Economics, 7(1), 99-124.

Ek, K., & Söderholm, P. (2010). The devil is in the details: Household electricity saving behavior and the role of information. Energy Policy, 38(3), 1578-1587.

Ellickson, R. C. (1994). The aim of order without law. Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft, 150(1), 97-100.

Ferguson, P. & Senie, D. (2000) Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing. RFC 2827, IETF BCP38

Ferraro, P. J., Miranda, J. J., & Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. American Economic Review, 101(3), 318-22.

Ferraro, P. J., & Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. Review of Economics and Statistics, 95(1), 64-73.

Festinger, L. (1954). A theory of social comparison processes. Human relations, 7(2), 117-140.

FICORA, 2015. Regulation on information security in telecommunications operations. [White paper]. Retrieved February 26, 2018 at https://www.viestintavirasto.fi/attachments/maaraykset/M67A_2015_EN.pdf

Florini, A. (2008). Making transparency work. Global Environmental Politics, 8(2), 14-16.

Fombrun, C. J., & Van Riel, C. B. (1997). The reputational landscape. Corporate reputation review, 1(2), 5-13.

Forsyth, D. R. (2000). Social comparison and influence in groups. In Handbook of Social Comparison (pp. 81-103). Springer, Boston, MA.

Financial Service Authority, (2008). Transparency as a regulatory tool. [White paper]. Retrieved March 19, 2018 at https://www.fca.org.uk/publication/discussion/fsa-dp08-03.pdf

Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *American Economic Review*, *94*(5), 1717-1722.

Gañán, C., Cetin, O., & van Eeten, M. (2015). An empirical analysis of ZeuS C&C lifetime. In Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (pp. 97-108). ACM.

Gal-Or, E. (1985). Information sharing in oligopoly. Econometrica: Journal of the Econometric Society, 329-343.

Gal-Or, E., & Ghose, A. (2005). The economic incentives for sharing security information. In Information Systems Research, 16(2), 186-208.

Ghose, A., & Rajan, U. (2006). The Economic Impact of Regulatory Information Disclosure on Information Security Investments, Competition, and Social Welfare. In WEIS.

Gordon, L. A., & Loeb, M. P. (2002). The economics of information security investment. ACM Transactions on Information and System Security (TISSEC), 5(4), 438-457.

Gordon, L. A., Loeb, M. P., & Lucyshyn, W. (2003). Sharing information on computer systems security: An economic analysis. Journal of Accounting and Public Policy, 22(6), 461-485.

Gottinger, H. W. (2003). Economies of network industries. Routledge.

Grønhøj, A., & Thøgersen, J. (2011). Feedback on household electricity consumption: learning and social influence processes. International Journal of Consumer Studies, 35(2), 138-145.

Gunningham, N., Grabosky, P., & Sinclair, D. (1999). Smart regulation: Designing environmental policy.

Gunningham, N., Kagan, R. A., & Thornton, D. (2004). Social license and environmental protection: why businesses go beyond compliance. Law & Social Inquiry, 29(2), 307-341.

Gupta, A. (2008). Transparency under scrutiny: Information disclosure in global environmental governance. Global Environmental Politics, 8(2), 1-7.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). Multivariate data analysis (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice hall.

Hardin, G (1968). "The Tragedy of the Commons" (PDF). Science. 162 (3859): 1243–1248.

Hawkinson, J. & Bates, T. (1996). Guidelines for creation, selection, and registration of an Autonomous System (AS). RFC 1930, IETF BCP6.

He, S., Lee, G. M., Han, S., & Whinston, A. B. (2016). How would information disclosure influence organizations' outbound spam volume? Evidence from a field experiment. Journal of Cybersecurity, 2(1), 99-118.

Helbing, D. (2013). Globally networked risks and how to respond. Nature, 497(7447), 51.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative health research, 15(9), 1277-1288.

Hutter, B. M., & Jones, C. J. (2007). From government to governance: External influences on business risk management. Regulation & Governance, 1(1), 27-45.

Huz, G., Bauer, S., & Beverly, R. (2015). Experience in using MTurk for Network Measurement. In Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data (pp. 27-32). ACM.

Internet Security Alliance, (2008). The Cyber Security Social Contract. Arlington.

Internet Society (2015). Addressing the challenge of IP spoofing. [white paper]. Retrieved January 18, 2018 at https://cdn.prod.internetsociety.org/wp-content/uploads/2017/08/ISOC-AntiSpoofing-20150909-en-2.pdf

ITU, (2017a). Global Cybersecurity Index (GCI). Geneva

ITU, (2017b). Measuring the information society. Volume 1 & 2. Geneva

Jhaveri, M. H., Cetin, O., Gañán, C., Moore, T., & Eeten, M. V. (2017). Abuse reporting and the fight against cybercrime. ACM Computing Surveys (CSUR), 49(4), 68.

Kahan, D. M. (2005). What's really wrong with shaming sanctions. Tex. L. Rev., 84, 2075.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. Journal of personality and social psychology, 65(4), 681.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

Kleinbaum, D. G. (1998). Survival Analysis, a Self-Learning Text. Biometrical Journal, 40(1), 107-108.

Krebs, B. (2016). The Democratization of Censorship. [Blog post]. Retrieved March 18, 2018 at https://krebsonsecurity.com/2016/09/the-democratization-of-censorship/

Kottler, S. (2018). February 28th DDoS Incident Report. [Blog post]. Retrieved March 18, 2018 at https://githubengineering.com/ddos-incident-report/

Krupka, E., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. Journal of Economic Psychology, 30(3), 307-320.

Kunreuther, H., & Heal, G. (2003). Interdependent security. Journal of risk and uncertainty, 26(2-3), 231-249.

Kührer, M., Hupperich, T., Rossow, C., & Holz, T. (2014a). Hell of a Handshake: Abusing TCP for Reflective Amplification DDoS Attacks. In WOOT.

Kührer, M., Hupperich, T., Rossow, C., & Holz, T. (2014b). Exit from Hell? Reducing the Impact of Amplification DDoS Attacks. In USENIX Security Symposium (pp. 111-125).

Lee, E. (2010). Information disclosure and environmental regulation: green lights and gray areas. Regulation & governance, 4(3), 303-328.

Li, F., Durumeric, Z., Czyz, J., Karami, M., Bailey, M., McCoy, D., ... & Paxson, V. (2016). You've Got Vulnerability: Exploring Effective Vulnerability Notifications. In USENIX Security Symposium (pp. 1033-1050).

Li, F., Ho, G., Kuan, E., Niu, Y., Ballard, L., Thomas, K., ... & Paxson, V. (2016). Remedying web hijacking: Notification effectiveness and webmaster comprehension. In Proceedings of the 25th International Conference on World Wide Web (pp. 1009-1019). International World Wide Web Conferences Steering Committee.

Lichtman, D., & Posner, E. (2006). Holding internet service providers accountable. Supreme Court Economic Review, 221-259.

Linden, L. L., Quarterman, J. S., Tang, Q., & Whinston, A. B. (2012). Reputation as Public Policy for Internet Security.

List, J. A., & Lucking-Reiley, D. (2002). The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign. Journal of Political Economy, 110(1), 215-233.

Lone, Q., Javed, M., Korczyński, M., Asghari, H, Luckie, M. & van Eeten, M. (in press). Using crowdsourcing marketplace for network measurement: the case of Spoofer.

Lone, Q., Luckie, M., Korczyński, M., & van Eeten, M. (2017). Using loops observed in traceroute to infer the ability to spoof. In International Conference on Passive and Active Network Measurement (pp. 229-241). Springer, Cham.

Lookabaugh, T., & Sicker, D. C. (2004). Security and Lock-in. In Economics of Information Security (pp. 225-246). Springer, Boston, MA.

Ludford, P. J., Cosley, D., Frankowski, D., & Terveen, L. (2004). Think different: increasing online community participation using uniqueness and group dissimilarity. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 631-638). ACM.

Macaulay, S. (1963). Non-contractual relations in business: A preliminary study. American sociological review, 55-67.

MANRS, 2015. Mutually Agreed Norms for Routing Security. [White paper]. Retrieved March 1, 2018 at https://www.manrs.org/wp-content/uploads/sites/14/2018/03/MANRS_PDF_Sep2016.pdf

Margetts, H., John, P., Escher, T., & Reissfelder, S. (2011). Social information and political participation on the internet: An experiment. European Political Science Review, 3(3), 321-344.

McAfee, (2018). The Economic Impact of Cybercrime - No Slowing Down. [white paper]. Retrieved March 12, 2018 at https://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime-summary.pdf

McConachie, A. (2014). Anti-Spoofing, BCP 38, and the Tragedy of the Commons. [Blog post] Retrieved February 20, 2018 at https://www.internetsociety.org/blog/2014/07/anti-spoofing-bcp-38-and-the-tragedy-of-the-commons/

McPherson, D., Baker, F. & Halpern, J. (2013). Source Address Validation Improvement (SAVI) Threat Scope. RFC 6959 IEFT

Morales, C. (2018). NETSCOUT Arbor Confirms 1.7 Tbps DDoS Attack; The Terabit Attack Era Is Upon Us [Blog post]. Retrieved March 18, 2018 at https://www.arbornetworks.com/blog/asert/netscout-arbor-confirms-1-7-tbps-ddos-attack-terabit-attack-era-upon-us/

Morris, R. T. (1985). A weakness in the 4.2 BSD Unix TCP/IP software. AT&T Bell Labs, Tech. Rep. Comput. Sci., 117.

Moore, T. (2010). The economics of cybersecurity: Principles and policy options. International Journal of Critical Infrastructure Protection, 3(3-4), 103-117.

Moore, T., & Clayton, R. (2008). The consequence of non-cooperation in the fight against phishing. In eCrime Researchers Summit, 2008 (pp. 1-14). IEEE.

Moore, T., & Clayton, R. (2011). The Impact of Public Information on Phishing Attack and Defense. Communications & Strategies, 1(81), 45-68.

Mulligan, D. K., & Schneider, F. B. (2011). Doctrine for cybersecurity. Daedalus, 140(4), 70-92.

Mulligan, D. K., Bamberger K. A. (2007) Security breach notification laws: views from chief security officers, Samuelson Law, Technology and Public Policy Clinic, University of California–Berkeley School of Law, Berkeley, California.

Open Resolver Project, (2013). [website]. Retrieved March 23, 2018 at http://openresolverproject.org

Patton, M. Q. (2002). Qualitative interviewing. *Qualitative research and evaluation methods*, *3*, 344-347.

Pawson, R. (2002). Evidence and policy and naming and shaming. Policy studies, 23(3), 211-230.

Postlewaite, A. (1998). The social basis of interdependent preferences. European economic review, 42(3-5), 779-800.

Powell, B. (2005). Is Cyberspace a Public Good-Evidence from the Financial Services Industry. JL Econ. & Pol'y, 1, 497.

Prince, M. (2014). Technical Details Behind a 400Gbps NTP Amplification DDoS Attack. [Blog post]. Retrieved March 18, 2018 at https://blog.cloudflare.com/technical-details-behind-a-400gbps-ntp-amplification-ddos-attack/

Quarterman, J.S. (2010). Internet Reputation Experiments for Better Security. [White paper]. Retrieved September 12, 2017, at https://labs.ripe.net/Members/jsq/internet-reputation-experiments-for-better-security

Quarterman, J. S., & Whinston, A. (2010). Fireeye's ozdok botnet takedown in spam blocklists and volume observed by iiar project. NANOG 48: Proceedings of the 48th North American Network Operators Group.

Quarterman, J., Sayin, S., & Whinston, A. (2011). Rustock botnet and asns. Telecommunications Policy Research Conference.

Romanosky, S., Telang, R., & Acquisti, A. (2011). Do data breach disclosure laws reduce identity theft?. Journal of Policy Analysis and Management, 30(2), 256-286.

Rossow, C. (2014). Amplification Hell: Revisiting Network Protocols for DDoS Abuse. In NDSS.

Roughan, M., Willinger, W., Maennel, O., Perouli, D., & Bush, R. (2011). 10 lessons from 10 years of measuring and modeling the internet's autonomous systems. *IEEE Journal on Selected Areas in Communications*, *29*(9), 1810-1821.

Saito, K. (2011). Role conflict and choice: Shame, temptation, and justifications. Working paper.

Schechter, S. E., & Smith, M. D. (2003). How much security is enough to stop a thief?. In International Conference on Financial Cryptography (pp. 122-137). Springer, Berlin, Heidelberg.

Schneier, B. (2004). Hacking the business climate for network security. Computer, 37(4), 87-89.

Shang, J., & Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, *119*(540), 1422-1439.

Shapiro, C. (1986). Exchange of cost information in oligopoly. The review of economic studies, 53(3), 433-446.

Shapiro, C., & Varian, H. R. (1998). Information rules: a strategic guide to the network economy. Harvard Business Press.

Shetty, N., Schwartz, G., Felegyhazi, M., & Walrand, J. (2010). Competitive cyber-insurance and internet security. In Economics of information security and privacy (pp. 229-247). Springer, Boston, MA.

Silverman, D. (2015). *Interpreting qualitative data*. Sage.

Stock, B., Pellegrino, G., Li, F., Backes, M., & Rossow, C. (2018). Didn't You Hear Me? - Towards More Successful Web Vulnerability Notifications.

Stock, B., Pellegrino, G., Rossow, C., Johns, M., & Backes, M. (2016). Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification. In USENIX Security Symposium (pp. 1015-1032).

Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect?. Current directions in psychological science, 11(5), 159-163.

Tang, Q., Linden, L., Quarterman, J. S., & Whinston, A. B. (2013). Improving internet security through social information and social comparison: A field quasi-experiment. WEIS 2013.

Tesser, A., Millar, M., & Moore, J. (1988). Some affective consequences of social comparison and reflection processes: The pain and pleasure of being close. Journal of personality and social psychology, 54(1), 49.

Thaler, R. H., & Sunstein, C. R. (1999). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT Yales University Press.

Thornton, D., Gunningham, N. A., & Kagan, R. A. (2005). General deterrence and corporate environmental behavior. Law & Policy, 27(2), 262-288.

van Eeten, M. J. and J. M. Bauer (2008), "Economics of Malware: Security Decisions, Incentives and Externalities", OECD Science, Technology and Industry Working Papers, 2008/1, OECD Publishing.

van Eeten, M., Bauer, J., Asghari, H., Tabatabaie, S., & Rand, D. (2010). The role of internet service providers in botnet mitigation an empirical analysis based on spam data.

van Erp, J. (2011). Naming and shaming in regulatory enforcement. In Explaining compliance: Business responses to regulation, 322.

Varian, H. (2000). Managing Online Security Risks. The New York Times. Retrieved February 26, 2018 at https://archive.nytimes.com/www.nytimes.com/library/financial/columns/060100econ-scene.html

Varian, H. (2004). System reliability and free riding. In Economics of information security (pp. 1-15). Springer, Boston, MA.

Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167-179.

Wheeler, L., & Miyake, K. (1992). Social comparison in everyday life. Journal of personality and social psychology, 62(5), 760.

Wills, T. A. (1981). Downward comparison principles in social psychology. Psychological bulletin, 90(2), 245.

Appendix 1:
Interpreting results of the Spoofer test

In this appendix, we introduce the process of analysis of the measurements of compliance with anti-spoofing filters collected via the Spoofer Project. In particular, we describe how test results collected at an IP address level are aggregated first to the IP prefix level and, subsequently to the entire autonomous system (AS). We provide examples in order to show and discuss the difficulties that might arise at every level during this process.

In Section 2.2.3, while discussing the best practices to deal with spoofed traffic, we mentioned some techniques used to measure compliance with such practices, devoting particular attention to the Spoofer Project. As you recall from that section, researchers of CAIDA developed a measurement application that volunteers can download and run to test the presence anti-spoofing filters on their network. The application tests the deployment of various types of filters, by attempting to send a sequence of packets with a spoofed source address. Each test is performed on the IPv4 address of the client and on the IPv6 address, if deployed. Once installed on a device, the application automatically runs both in the background once a week and every time the device is connected to a new network.

As you recall, the test can have one of the following outcomes:

- *received*: the spoofed packet was received, which means that source network does not implement ingress filtering;
- *rewritten*: the spoofed packet was received but the original source address was changed en-route, which indicates the presence of a Network Address Translation (NAT) that rewrites the header of the spoofed packet;
- *blocked*: the spoofed UDP packet was not received, but the TCP packet (unspoofed) was, meaning that the spoofed packet was dropped by an in-network filter;
- *unknown*: neither spoofed nor unspoofed packet was received.

Tests results are collected and displayed online on CAIDA's website, with different level of aggregations. For each test, the following information are recorded (as shown in Figure 21):

- *Session & timestamp:*

A session ID is assigned to each test, together with the timestamp of when the test has been performed

- *Client IP block*

The IP address of the client performing the test. The precise IP addresses are anonymised for security concern, and instead the IP prefix (range of addresses) is shown. In case of IPv4 (32-bit addresses) the last 8 bits are masked (resulting in a /24 prefix containing 256 addresses), whereas in case of IPv6 (128-bit addresses) the last 88 bits are masked (resulting in a /40 prefix).

- *ASN*

As you recall, an autonomous system (AS) is, loosely speaking, a portion of the Internet under a single administrative control. Every autonomous system is assigned a unique autonomous system number (ASN), fundamental to identify that network on the Internet.

- *Country*

The country in from the test is done.

- *NAT*

This field indicates whether the divide doing the test is behind a network address translation (NAT). As you recall, the presence of a NAT might interfere with the measurement, as it might rewrite the original spoofed address with the address of the NAT device.

- *Spoof Private*

As the application tests the presence of various type of filters, a further distinction is made on the base of the type of address the application tries to spoof: private or routable addresses. Private addresses are commonly used in private environments (e.g. homes, enterprises and LANs), and must not be propagated outside these private networks. Routable addresses, instead, are "public" IP addresses, that are be assigned on the public core of the Internet. This field indicates the result of the test done by trying to spoof a private address.

- *Spoof routable*

Similarly, this field indicates the result of the test done by trying to spoof a routable address.

- *Adjacency spoofing*

The application also tries to establish the granularity of any filtering, by incrementally spoofing addresses in adjacent prefixes. This "neighbour spoof" attempts successively larger boundaries, until spoofing an address in an adjacent /8

- *Results*

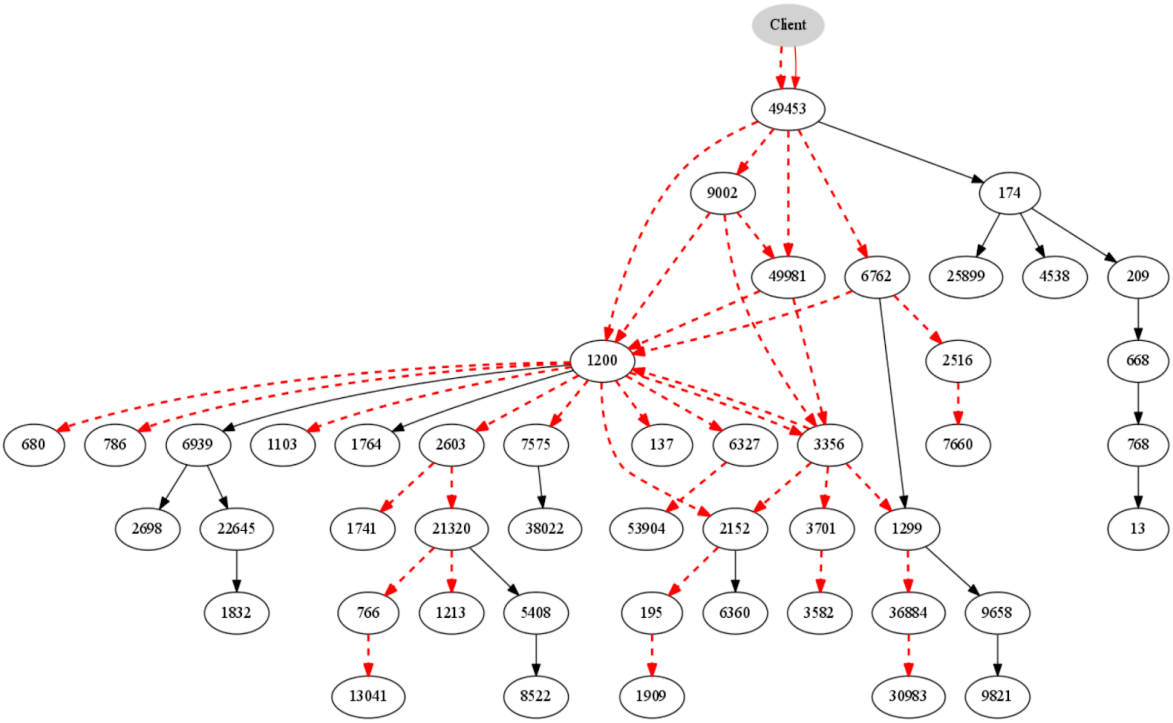
A report of the test is also created, providing additional information on the test such as the AS travelled by the test packet (Figure 22).

Figure 20. Tests collected from the prefix 213.152.165.0/24

Session	Timestamp (UTC)	Client IP Block	ASN	Country	NAT	Spoof Private	Spoof Routable	Adjacency Spoofing	Results
342260	2017-10-27 01:46:00	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
342258	2017-10-27 01:45:35	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
322506	2017-09-24 15:33:56	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
319436	2017-09-19 16:36:53	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
314692	2017-09-12 15:30:30	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
310138	2017-09-05 14:24:09	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
305892	2017-08-29 13:07:25	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
301441	2017-08-22 12:00:32	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report
299620	2017-08-19 03:54:34	213.152.165.x/24	49453 (GLOBALLAYER)	nld (Netherlands)	no	blocked	received	/8	Report

Figure 21. Details of test results.

The full red line marks the path of the packet with a spoofed private address (blocked by AS49453), the dashed red line represents the path of the packet with a spoofed routable address (not blocked by AS49453).



So far, we have discussed the results of single tests. But how can we generalise single test results to the level of the IP prefix and, subsequently, to the entire autonomous system?

In the example provided in Figure 21, it is quite straightforward to conclude that the prefix 213.152.165.0/24 blocks packets with a spoofed private source address but not those with a spoofed routable address. However, this is not always the case. Consider, for example, the case illustrated in Figure 23, that shows all the results from the prefix 195.8.192.0/24. In this case, some tests are positive (i.e. reveal the lack of anti-spoofing filters) and some are negative (i.e. the spoofed packets are blocked). This example reveals some of the limitation of this type of measurements: relying on volunteers provide only a partial view on the state of the prefix. Viewing Figure 23 in its entirety, we can conclude that the prefix does not correctly deploy anti-spoofing filters. However, if we had at disposition only the test between 13-7 and 18-7 our conclusion would have been that filters were deployed! (Note that in this example the time interval in which tests are negative consists of only 5 days, but in other cases this interval can be much longer). Moreover, if our time windows instead were between 12-7 and 18-7 we would have seen the first positive test followed by negative tests, and we could have inaccurately concluded that the operator remediated by deploying anti-spoofing filters. Finally, note that the last test (positive) was performed at the end of February 2018. In other cases, the last test revealing the lack of filtering is several months older, and it might be that operators have since remediated, but no additional tests were performed.

We can make some speculations on the reason of such heterogeneity in the test results. A first cause can be the fact that results are aggregated at a IP prefix. As mentioned, a /24 prefix include 256 addresses (more precisely, 254 available host, as 2 addresses are reserved as network base address and broadcast address). It might be that these 254 addresses are assigned to end users in different subnetworks, and that some intermediary piece of routing equipment in a subnet implement anti-spoofing filtering while other do not. A second option is that the network operators enable and disable filters depending on other factors. As you recall from Section 2.2.3, uRPF, the automatic way to filter spoofed traffic, may lead to significant drops in performances of the network infrastructure. Thus, it might be that, in case of large volume of traffic on that particular interface, operators disable uRPF. Finally, it is important to mention that the Spoofer application is also often used by operators to check routing configurations. Therefore, the heterogeneity in the test results can be due to modifications to these configurations. Nevertheless, what precisely causes similar situations remains unknown.

Figure 22. Test collected from the prefix 195.8.192.0/24.

Session	Timestamp (UTC)	Client IP Block	ASN	Country	NAT	Spoof Private	Spoof Routable	Adjacency Spoofing	Results
417182	2018-02-26 13:33:05	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	received	received	/8	Report
375873	2017-12-19 12:31:58	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	received	received	/17	Report
357587	2017-11-20 19:43:32	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	received	/31	Report
284329	2017-07-28 11:17:03	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	received	received	/8	Report
284079	2017-07-28 03:16:22	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	blocked	none	Report
283386	2017-07-27 09:44:56	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	received	received	none	Report
		2a00:19d8:xx::/40	42004 (ULGRP-AS)		no	received	received	/16	
276963	2017-07-19 12:03:44	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	received	received	none	Report
275843	2017-07-18 05:53:51	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	blocked	/28	Report
		2a00:19d8:xx::/40	42004 (ULGRP-AS)		no	blocked	blocked	/64	
275812	2017-07-18 04:55:45	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	blocked	/28	Report
		2a00:19d8:xx::/40	42004 (ULGRP-AS)		no	blocked	blocked	/64	
275761	2017-07-18 03:07:03	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	unknown	unknown	none	Report
271900	2017-07-13 08:37:31	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	blocked	/29	Report
271739	2017-07-13 04:17:56	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	blocked	/28	Report
		2a00:19d8:xx::/40	42004 (ULGRP-AS)		no	blocked	blocked	none	
271719	2017-07-13 03:32:48	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	blocked	blocked	/28	Report
		2a00:19d8:xx::/40	42004 (ULGRP-AS)		no	blocked	blocked	/64	
271024	2017-07-12 10:35:03	195.8.192.x/24	42004 (ULGRP-AS)	gbr (United Kingdom)	yes	received	received	/8	Report
		2a00:19d8:xx::/40	42004 (ULGRP-AS)		no	received	received	/16	

Figure 23. Test collected from AS20860.

IP block	Latest test	Spoof Private	Spoof Routable
78.129.250.x/24	2017-01-24 09:12:29 GMT	received	received
87.117.234.x/24	2017-05-01 21:12:02 GMT	received	received
130.185.150.x/24	2016-02-29 15:02:22 GMT	received	received
217.147.89.x/24	2016-11-22 17:52:30 GMT	received	received
5.62.14.x/24	2017-03-11 16:18:03 GMT	unknown	received
78.129.171.x/24	2017-03-19 22:00:07 GMT	blocked	blocked
81.94.192.x/24	2016-04-26 21:14:36 GMT	blocked	blocked
87.117.197.x/24	2016-03-21 19:29:59 GMT	blocked	blocked

Similar problems occur when considering the situation to a AS level. Figure 24 shows the results of the most recent tests from some prefixes of AS20860 (IOMART-AS). In this case, some prefixes allow spoofing on both private and routable addresses, one prefix seems to block only private addresses, and three prefixes appear to have anti-spoofing correctly deployed.

As a last level of aggregation, ASes are grouped by country. Also in this case, problems might arise, as ASes may span over multiple countries. For example, AS174 (COGENT-174), presents tests from the Netherlands, the US, Italy, the UK, Czech Republic and Canada. When looked up in the WHOIS database, AS174 appears to be registered in the US. However, when the IP prefixes of the test client are looked up, these (sometimes) appears to be assigned in other countries. Again, it is difficult to establish the cause of such mismatch.

In light of what has been discussed so far, what metrics can be used to generalise single test results to the entire IP prefix? First, we restrict the outcome of the test by not considering "unknown" results, as they do not carry any information. Then, we propose the following metrics:

- *Spoofable*: an IP prefix from which the most recent test is "received";
- *Mixed*: an IP prefix from which at least one test is "received", but the most recent test is not;
- *NAT-blocking*: an IP prefix from which the majority of test is "rewritten" (and no "received" test);
- *UnSpoofable*: a prefix from which the majority of test is "blocked" (and no "received" test).

Note that in our metrics we do not distinguish between private and routable spoofing, as BCP38 prescribes to filter both type of addresses. Also, note that the distinction between "spoofable" and "mixed" is functional to leave the benefit of the doubt to operators whose networks do not show consistent evidences of spoofing. In fact, incorrectly categorise these networks as "spoofable", might have a negative effect on out intervention, as a high number of false positives might reduce the reliability of the measurements and our credibility as discloser. Though this distinction is used in the information we disclose, during the stage of analysis we will consider both "spoofable" and "mixed" as if they are not deploying anti-spoofing.

A similar reasoning applies to extend the metrics just elaborated to the case of ASes. Specifically, we will consider:

- *Spoofable*: an AS with at least one "spoofable" IP prefix;
- *Mixed*: an AS with more "mixed" IP prefixes than "spoofable" IP prefixes;
- *NAT-blocking*: an AS with a majority of "NAT-blocking" IP prefixes (and no "spoofable" prefixes);
- *UnSpoofable*: an AS with a majority of "UnSpoofable" IP prefixes (and no "spoofable" prefixes).

Finally, we also define remediation on the base of the results of the test. Figure25 shows an example of one prefix that presents evidences of remediation. We can observe a series of positive tests (in red), followed by a sequence of negative tests. This suggests that the operator deployed ingress filtering

between the 6-7-2017 and the 8-11-2017 on this prefix. Nevertheless, it is interesting to notice that, when we looked up all the other prefixes of AS9268 tested, we noticed the presence of tests that revealed the lack of anti-spoofing on other subnets after the 8-11-2017, implying that remediation was only limited to the particular prefix shown in Figure 25.

Figure 24. Evidence of remediation from the prefix 180.214.68.0/24.

Session	Timestamp (UTC)	Client IP Block	ASN	Country	NAT	Spoof Private	Spoof Routable	Adjacency Spoofing	Results
374959	2017-12-17 23:37:04	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	blocked	blocked	none	Report
372429	2017-12-13 23:57:17	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	unknown	unknown	none	Report
358920	2017-11-23 02:58:02	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	blocked	blocked	none	Report
357099	2017-11-20 01:09:00	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	blocked	blocked	none	Report
349669	2017-11-08 03:46:28	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	blocked	blocked	none	Report
265959	2017-07-06 00:12:34	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	unknown	received	/8	Report
262716	2017-07-02 03:44:40	180.214.68.x/24	9268 (OVERTHEWIRE-AS-AP)	aus (Australia)	yes	unknown	received	/9	Report

Appendix 2: Preliminary interviews

In order to gain a more practical insight into the problem of IP spoofing and the diffusion of SAV, a sample composed of experts in network management and Internet governance has been invited to participate in the research. Six semi-structured interviews and three additional informal interviews have been conducted, to investigate what are ISPs' incentives to deploy SAV and which factors have a larger influence on these incentives. In this Appendix, we take a closer look at the methodology used to conduct the interviews and to analyse the results. First, the aim of the interviews is discussed in Section 1. Secondly, the methodology is described in Section 2. The analysis of results is presented in Section 3, followed by a discussion of the limitations in Section 4.

1. Aim of the interviews

The objective of the interviews is to investigate ISPs' incentives to deploy SAV on their network. Firstly, we seek to gain additional knowledge on the problem of IP spoofing and, in particular, on its perception from ISPs' point of view. Secondly, our aim is to get a technical insight into the deployment of SAV, to study the costs and benefits associated with its implementation and its possible technical limitations. Finally, the interviews will help us to validate findings of previous research about the factors that can determine ISPs' incentives.

2. Methodology

This section describes the methodology adopted to conduct the interviews. To begin with, the general approach of the interviews and their structure are explained in Section 2.1. Then, Section 2.2 reports the interview guide, containing the protocol and the questions asked to participants. Lastly, Section 2.3 discusses the composition of the interview sample.

2.1 Approach

When developing a framework to conduct interviews, an important trade-off to deal with refers to the extent to which the interview is structured. Structured (or formal) interviews have a rigid framework, in which the interviewer poses the same predefined and standardised sequence of questions (including follow-up questions) to each participant (Patton, 2002). In this way, the interview is approached in a neutral, systematic manner, with no space for the interviewer to improvise or to deviate from the interview guide containing the questions (Silverman, 2015). Conversely, unstructured (or informal) interviews contain open-ended questions, without necessarily a prefixed order. This type of interview provides the interviewer with complete flexibility to add or skip questions as the discussion progresses (Silverman, 2015).

In order to touch a number of prearranged topics, and at the same time to have the flexibility to explore new topics that emerge during the discussion, we opted for conducting semi-structured interviews. Semi-structured interviews combine the systematic approach of structured interviews (i.e. the predefined list of questions and topics to discuss) together with the elasticity of unstructured interviews (i.e. the possibility to work flexibly with the guide) (Silverman, 2015).

An interview guide has been developed, based on previous research, to cover the main factors that can relate to ISPs' incentives. Moreover, the guide included hints to steer the conversation, and to stimulate it in the case participants hesitate to respond. The guide consisted of 12 questions, divided into in four different parts of the interview: an introduction (PART I), a part on the factors that can prevent ISPs from implementing SAV (PART II), one on the strategies to incentivise ISPs to deploy SAV (PART III), and a conclusion (PART IV). The two central parts of the interview consist of 8 questions, and represent the core of the interviews, aimed at investigating ISPs' incentives to deploy SAV. As the theme of ISPs' incentives has already been quite debated in the literature (in part, specifically for the case of IP spoofing, but also in general for the role of ISPs in other security issues, such as botnets, spam, phishing or hosting malicious contents), we designed each question of this central part to probe one particular factor related to ISPs' incentives. In particular, seven main factors were investigated: *awareness, technical limitations, costs, benefits, information on non-compliant ISPs, reputation/peer pressure and liability.*
Most of the questions were fairly open-ended, in order to let participants express their general opinion first. Eventually, the hints in the guide were used to formulate more specific questions, in case answers were vague or unfocused.

After introducing the general topic of the research, participants were asked for the permission to tape-record the conversation. As no participant objected, all interviews have been recorded and transcribed verbatim.

An introductory, open question was used to start the conversation:

Could you tell me about your expertise with IP spoofing and BCP38?

The central part of the interview was designed to discuss factors that can prevent ISP from implementing SAV. A preliminarily list of factors has been gathered by considering the academic literature on the problem, as well as articles and blog posts retrieved from the Internet. These factors included the level of awareness, costs and (lack of) benefits, and possible technical limitations. Moreover, we questioned to what extent the lack of transparency about which operators do/do not implement SAV affects its diffusion. As the disclosure of this information is central in our research, the interviews will help us to understand its relevance, before releasing it. Lastly, participants were asked about the role of peer pressure and liability in ISPs' incentives to implement SAV.
Despite the questions were organised according to a logical order in the guide, the succession of the questions during the interviews depended on participants' answers. For instance, when a participant was asked about the costs of implementing SAV, his response focused on the possible reputation damage related to the lack of SAV. Subsequently, questions about the role of reputation, peer pressure and

liability were asked. After discussing these themes, the remaining questions were asked. Generally, all the questions have been asked to every participant.
In order to explore also themes not included in the guide, probing questions, like: "Can you think to other factors that can contribute to the decision to implement SAV?" were asked in multiple occasions.

After all the questions were posed, participants were also asked if they had any final remarks or comments to add. To conclude the interview, participants were thanked for their collaboration.

On top of the semi-structured interview described above, three informal meetings with participants that had no time to conduct the whole interview, or whose expertise was limited. These informal interviews have turned useful to discuss the problem to a general level, quickly going through the interview questions. Moreover, during these meetings, we asked for contact information of more appropriate the people to conduct the full interview with.

2.2 Interview guide

PART I: introduction

- Provide background of the research;
- Discuss general structure of the interview;
- Explain results will be anonymised;
- Ask permission to record.

Could you tell me about your of expertise with IP spoofing and BCP38?

PART II: factors that steer implementation of SAV

What do you think is the level of awareness about the relevance of BCP38 among ISPs, regulatory organisations, and end users?

- Which actors need to be made more aware?
- Do you think the moderate adoption of BCP38 is a problem of lack of awareness?
- Do you think there are other factors that can explain the moderate adoption of BCP38?
- How often the issue of compliance with BCP38 comes out, and in which occasions?

To what extent technological limitations may affect the implementation of BCP38?

- In terms of difficulties to implement/lack of know how
- In terms of compatibility with other services/future network enlargement

What are the costs of implementing BCP38?

- What are direct costs (new equipment/installation…)?
- What are indirect costs (maintenance/training of personnel…)?

What type of benefits does BCP38 brings to the ISP that implements it?

- Can it be used as sign of trustworthy/good reputation?

Do you think that ISPs with a smaller customer base may perceive these factors differently?

PART III: strategies to incentivise ISPs to deploy SAV

By looking at the traffic on the network (or to other source of data), how feasible is it to detect non-compliant ISPs?
Is there knowledge about which operators are compliant with BCP38 and which are not?

- How do you use information about non-compliant ISPs?
- Do you send or receive notifications about non-compliant ISPs?
- From who? How do you act on such notifications?

What effects do you think that releasing information about non-compliant ISPs would have on the diffusion of BCP38?

Do you know if there is any liability for not deploying BCP38?

- What is the responsibility of the ISP in which an attack using a spoofed address originates towards the victim?

PART IV: conclusion

Which of factors we discussed do you think have more importance for ISPs in the decision to implement BCP38?

- Is there something that we have not discussed, but you think may be relevant to my research?
- Do you perhaps know other people who can contribute to this research as well?

- stop recording;
- give full context of the research;
- thanks participant for his collaboration;

2.3 Sample

A convenience sample composed of experts in network management and Internet governance has been invited to participate in the research, in order to gain additional knowledge about the diffusion of SAV from a practical perspective.

The sample includes personnel of a major Dutch network operators, as well as and experts involved in the governance of Internet security. Firstly, we contacted personnel of different department of KPN N.V., the main ISP in the Netherlands, to discuss costs, benefits and limitations of the deployment of SAV. In addition, we contacted members of RIPE, of the European NOG and of a major Dutch Internet Exchange Point. As for the regulatory side, members of Internet Society, of the European CERT, and the ENISA (European Network and Information Security Agency) have been invited, in order to acquire information on the strategies that have been pursued to prompt operators to deploy SAV, and on the limitations of such strategies. Despite some invitations have been declined, and several got no reply at all, we managed to gather a sample composed of 6 experts:

3 belonging to the network engineering department of KPN;
1 from KPN Security Operation Centre;
1 from KPN Cert;
1 from the Internet Society.

3. Analysis

This section reports the results of the interviews. First, in Section 3.1, the process of interpretation and analysis of the interview is described. Subsequently, in Section 3.2, the results are presented.

3.1 Interpretation

The process of analysis of the interviews, from the recording to the final results, has been approached with a fairly deductive perspective. In fact, the interviews have been designed considering various themes provided by the literature, in order to validate and extend these findings. Since the interview questions are already pretty structured and divided by theme, a top-down approach (from themes to quotes) was preferred in the analysis of the results. In particular, we opted for using Direct Content Analysis (Hsieh & Shannon, 2005). In this type of analysis, prior research is used to identify key themes or variables as initial coding categories. We started with a list of 7 codes, prepared in the design phase of the interviews (these were the main factors based on past research):

1. *Awareness*
2. *Technical limitations*
3. *Costs*
4. *Benefits*
5. *Information on non-compliant ISPs*
6. *Reputation/peer pressure*
7. *Liability*

To analyse the responses, the transcripts of the interviews have been broken down into short paragraphs of 2/3 sentences, and each short paragraph has been assigned to one of the above codes. It is interesting to notice that, despite questions were already arranged around these categories, different paragraphs of the same answers often referred to multiple codes. This happened because the factors we question are often interconnected, and can influence each other (e.g. technical limitations of old equipment can be translated in the cost of acquiring new equipment).

In the following subsection, we are going to discuss the results of the interview grouped by code.

3.2 Results

Awareness

Participants generally agreed on claiming that the level of awareness about IP spoofing is high, at least in the western world and among larger operators. Two participants, however, highlighted that IP spoofing itself is not a major threat to ISPs, and that anti-spoofing is mainly done as a part of anti-

DDoS. One participant said: *"Spoofing is a general problem for the Internet, but not a special focus for ISPs.* […] *It is in the pre-attack leg, and spoofed traffic do not cause problems to anyone. The problem is with reflection and amplification DDoS, and the malicious traffic in these attacks is from, let's say, a DNS server, which is not spoofed.* […] *And that is legitimate traffic from a legitimate source.* […] *And also if you mitigate spoofing, DDoS attacks are still possible, less volumetric and maybe more traceable, but still dangerous"*. Another participant added: *"When it comes to you and me as ISPs, we don't care too much, because spoofed traffic is not a problem…it's even less than spam, and it is not directly harmful for us* […] *but you do it because it's the right thing to do!"*.

"We keep seeing massive reflection and amplification DDoS attacks based on spoofing" said a participant, *"so at least from a technical side, we are aware about it and we know what is the importance to fight it"*. The problem, he suggested, may lie in the gap between this technical side and the level where the decision to do anti-spoofing is taken. In particular, two participants claimed that the business side of ISPs should made more aware about the importance of implementing SAV. Another participant argued that the problem of IP spoofing stands from a broader problem of awareness about the importance of similar best practices (referred to as *"network hygiene"*).

Moreover, most of the participant acknowledge the global scale of the problem, suggesting how this aggravate the situation. One participant observed that in the Netherlands the community is relative aware about SAV, but added: *"The community is local, while the problem is global"*. In addition, many participants share the common view that: *"There are countries like China or Russia in which operators simply do not care"*. Consequently, some participant speculated that some ISPs might wonder *"What can be my contribution, when China is the bigger polluter?"*.

Lastly, two participants mentioned that the problem of IP spoofing is sometimes discussed in different occasions like regional and international meetings and conferences. However, the people attending these meetings are often those *"More actively engaged in doing proper security and management of their environment, and it is quite easy they are aware about BCP38, also because it has been around for a while"*. Evidently, the challenge is diffuse awareness outside this group.

Technical limitations

In general, technical limitations may arise in the process of implementing BCP38, but the type and the impact of these limitations depend on the specific case.

For consumer connections and access points, experts generally agreed that limitations are marginal:
Edge configuration are not difficult to set up, and many protocols have built in security function to prevent spoofing by default. Three participants identified slightly more problematic those situation in which customers host their own IP space, frequently request additional ranges, or have a fast-growing network because of the administrative hassle of modifying configurations.
For very large networks, involving also small and medium enterprises, limitations are heavier, and the implementation of SAV requires more ad hoc configurations:
In case of multi-homed networks, BCP38 has been updated to BCP84, that discusses the use of Unicast Reverse Path Forwarding (uRPF). This is a function that operators can switch on to modify the routing

policies to do SAV. One participant, however, claimed that the risk of cutting off customers with uRPF increase dramatically, and said he heard of drops in performances of 30%, which force operators to switch it off.

On top of this, two participants have also suggested that the discriminant is in the type of configuration: automatic or manual. Manual configurations are more error prone, increase fragility of the network and make debug and maintenance more complex.

Moreover, two participants highlighted that the implementation of BCP38, despite not difficult, is a process that requires time for setting up and testing. Without considering the time to take the decision, the implementation itself may last anywhere between few weeks and several months, depending on the complexity of the network.

Two participants noted that small, regional ISPs can experience more difficulties in the implementation and in the testing phase, due to lack of skilled personnel. However, these operators generally do not deal with large business clients, and therefore face only moderate limitations.

Finally, while arguing about the difficulty of implementing SAV, two participants mentioned that: *"If ISPs can set up BGP peering, then they can also do the basic anti-spoofing".*

Costs

Experts agree that costs of implementation of SAV are fairly low. No new equipment is needed, and the majority of router today support automatic configurations (there might be exception, for examples, old lines that do not support uRPF). One participant pointed out that the costs of deploying and maintaining SAV are *"Part of ISPs' daily job as a maintainer of the network…to have good hygiene on your BGP configuration"*. Three participants, however, claim that the administration and the continuous update of filtering list can require particular attention, and that it definitely represents a cost (especially for smaller operators). The cost of training of personnel, instead, is minimal, and according to one participant, is no more than few slides in the configuration manual.

In addition, three participants suggested that more than a cost, *"Introducing any new feature is a risk, because the technology can fail, and then the cost of cutting off customer is very high"*. This view is shared by another participant, who argued *"If something work, you don't want to touch it, especially when you don't have a direct profit".*

Benefits

Without doubt, the main benefit of SAV is to *"prevent customers from being used as launch-pad for reflection and amplification DDoS attack"*. According to one participant, this can be seen as reputational advantage (towards customers, and towards the community), or as compliance to regulations (in countries in which anti-spoofing is mandatory, e.g. Finland). However, participants were able to identify other, less direct benefits:

- *Avoid spoofed management access to network stations* (1 participant);
- *By dropping non-authenticated packets, the deployer ISP saves resources* (2 participants);

- *Maintain the core of the network cleaner and improve reliability of traffic analysis* (1 participant);
- *Trust and good reputation* (5 participants).

Information on non-compliant ISPs

According to most of participants, reliable information about which ISPs are compliant with BCP38 and which are is not feasible to obtain. One participant said that: *"Operators have a picture of their peers, who are applying certain hygiene and who do not, and I think that spills over to other things as well, like incorrect routing announcement, spoofed traffic of hosting some questionable clients. […] But I don't think operators specifically look at spoofing"*. Another participant claimed that ISPs do not look for that piece of information, expect *"When there is an incident, and you try to get in contact with the upstream provider to shut down the traffic coming from that ISP"*.
It seems that information about non-compliant ISPs, though can be relevant in regulatory context, is not very important for defenders.

Next, two participants argued that information about non-compliant ISPs might be deducted from the attacks, with the support of organisations like RIPE and IANA that manage IP addresses, but not without difficulties. Other two claimed that the only way to measure if a network is compliant is by trying to spoof IP addresses from that network. One of these mentioned CAIDA's Spoofer Project.

Moreover, some participants mentioned initiative like the Trusted Network Initiative, DCB (Dutch Continuity Board), MANRS (Mutually Agreed Norms for Routing Security) aimed at incentivising ISPs to implement a series of best practices to mitigate DDoS, including BCP38. The core idea that BCP38 compliance is a sign of trust and responsibility. One participant, in particular, suggested that ISPs that implement similar practices should promote it and show their compliance, for example on their peering policies.

In addition, one participant mentioned two initiatives, OIRTO (Operational Incident Response Team Overleg) and OPS-Trust, active in the field of incident response. These initiatives operate via mailing lists, and send notifications to the source networks of DDoS attack to instigate clean-up of compromised resources.

Reputation/peer pressure

Reputation has been named multiple times during the interviews. Most of participants mentioned good reputation as a benefit of implementing SAV, arguing that maintaining their status is an important incentive for ISPs. In addition, three participants claimed that reputation and peer pressure are the most effective incentives for operators to deploy SAV. When asked about the effect of releasing information about non-compliant operators, they argued that it is a viable way to incentivise ISPs. However, two of them also explained the limit of this approach: *"Sure, with the assumption people care about reputation…"*. Another one argued that reputation might play a role in the negotiation of peering agreements between two networks: *"If you don't do anti-spoofing, I will not peer with you"*. However, ISPs might not have this sort of leverage, as they need peering to reach global connectivity.
This shows a bigger problem, better explained by another participant: *"First, you have to establish a norm (in this case IP spoofing is not acceptable). Then, you can start to create small communities with*

focused peer pressure". However, the same participant suggested that, since IP spoofing is not considered as a critical problem for ISPs, also peer pressure will have limited results.
In addition, two participants suggested that the size of an ISP can mediate the effect of this incentive. One said that: *"The bigger you are, the more you are concern about reputation and performance limitations"*.

Liability

When discussing the issue of liability, two participants claimed that making SAV mandatory would be the most effective way to incentivise ISPs. Conversely, three participants also argued that imposing liability is not a feasible solution. In particular, one participant mentioned the case of Finland, where regulations prescribe ISPs to do anti-spoofing. The same participant then explained that this type of regulations is relatively difficult to enforce, given the problems in testing compliance, and are expensive to maintain, a view shared by another other participant. In addition, one participant noticed that in reflection and amplification DDoS the attack is often sent by devices in a botnet. This relates to an ongoing debate on the role of ISPs in the mitigation of botnets, an issue with a much broader scope that this work.

Summary

Awareness:

- generally high (especially among large IPSs) (5 participants)
- IP spoofing is not a threat for ISPs, the problem is DDoS (2 participants)
- broader problem of awareness about network hygiene
- global scale of the problem "China and Russia" are the biggest polluter (4 participants)

Technical limitations:

- no problems for access provider (except some very particular cases) (5 participants)
- more problematic with SME or business users, see BCP84 and uRPF (3 participants)
- manual or automatic configuration? Manual is error prone, increase fragility and makes testing and maintenance more complicate (2 participants)
- time: from few weeks to several months, depending on type of network (2 participants)
- difficulty: "if you can set up BGP peering, you can also set up BCP38" (2 participants)

Costs:

- pretty low (4 participants):
    - no new equipment (except when infrastructure is old)
    - maintenance requires attention (3 participants) but is part of your daily job to keep the network clean (1 participant)
    - more than a cost, is a risk (3 participants): "you don't want to cut off customer, especially for functionality that are not profitable"

Benefits:

- Avoid spoofed management access to network stations (1 participant);
- By dropping non authenticated packets, the deployer ISP saves resources (2 participants);
- Maintain the core of the network cleaner and improve reliability of traffic analysis (1 participant);
- Trust and good reputation (5 participants).

Information about non-deployer ISPs:

- very difficult to obtain (4 participants)
- useful in regulatory context, not for defenders (2 participants)
- from attacks data, but requires high coordination (2 participants)
- the only way is to try to spoof from within the network (2 participants), 1 mention CAIDA

Reputation:

- maintaining status is definitely important for ISPs (4 participants)
- peer pressure will not work because IP spoofing is not perceived as a problem (1 participants)
    - first, norm setting: "IP spoofing must not be tolerated"
    - then, create small communities with focused peer pressure

Liability:

- the most effective way (2 participants)
- not doable (3 participants):
    - difficult and expensive to enforce, difficulties in measuring compliance
- related to ISPs' responsibility in botnet, ongoing debate (1 participants)

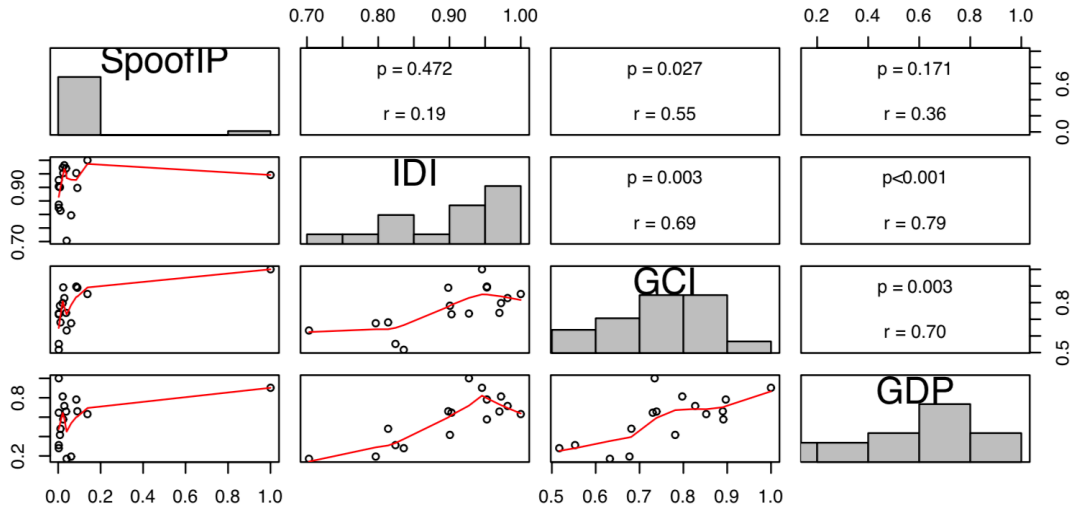
4. Limitations

In this final section, we discuss the limitations affecting our interview methodology.
First and foremost, the structure of the interview and of the questions has been based on prior research. From the literature, a list of factors to discuss during the interview has been prepared. The use of existing theories presents some inherent limitations, since the interviewer approaches the discussion with an informed, but nevertheless biased viewpoint. Therefore, it is more likely that the evidences found are fairly supportive, instead of non-supportive of a theory (Hsieh, Shannon 2005). Moreover, the probing questions might lead some participants to response in a way that please the researcher.
A second major limitation is related to the sample size and its composition. Despite the invitation to participate has been sent to 7 different organisations, only two took part to the interviews. The majority of the participants came from a technical department of KPN. This has shed a partial light on ISPs' incentives to deploy SAV, from the point of view of a large, established ISP that implement SAV. To address this issue, future research should focus on non-deployer ISPs as well.

Appendix 3:
Cluster analysis

```
> # dataset normalised
> mydata
        SpoofIP       IDI       GCI       GDP
AUS 0.085106383 0.9526012 0.8966268 0.7818108
BEL 0.002659574 0.9028902 0.7301415 0.6457048
CAN 0.090425532 0.8982659 0.8900979 0.6601422
FRA 0.023936170 0.9526012 0.8911861 0.5771038
DEU 0.037234043 0.9699422 0.7388466 0.6587642
GRC 0.002659574 0.8358381 0.5168662 0.2807616
IRL 0.002659574 0.9271676 0.7344940 1.0000000
ITA 0.010638298 0.8138728 0.6811752 0.4803326
NLD 0.029255319 0.9815029 0.8269859 0.7151358
POL 0.061170213 0.7965318 0.6768226 0.1944975
PRT 0.002659574 0.8242775 0.5527748 0.3106699
ESP 0.007978723 0.9005780 0.7812840 0.4171495
SWE 0.021276596 0.9722543 0.7976061 0.8133632
TUR 0.039893617 0.7028902 0.6322089 0.1700855
GBR 0.138297872 1.0000000 0.8520131 0.6316902
USA 1.000000000 0.9456647 1.0000000 0.9025399

> # descriptive statistics
> summary(mydata)
    SpoofIP             IDI              GCI              GDP        
 Min.   :0.002660   Min.   :0.7029   Min.   :0.5169   Min.   :0.1701  
 1st Qu.:0.006649   1st Qu.:0.8329   1st Qu.:0.6801   1st Qu.:0.3905  
 Median :0.026596   Median :0.9150   Median :0.7601   Median :0.6387  
 Mean   :0.097241   Mean   :0.8986   Mean   :0.7624   Mean   :0.5775  
 3rd Qu.:0.067154   3rd Qu.:0.9569   3rd Qu.:0.8615   3rd Qu.:0.7318  
 Max.   :1.000000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
```

Figure 25. Scatter plot, correlation matrix and histogram.



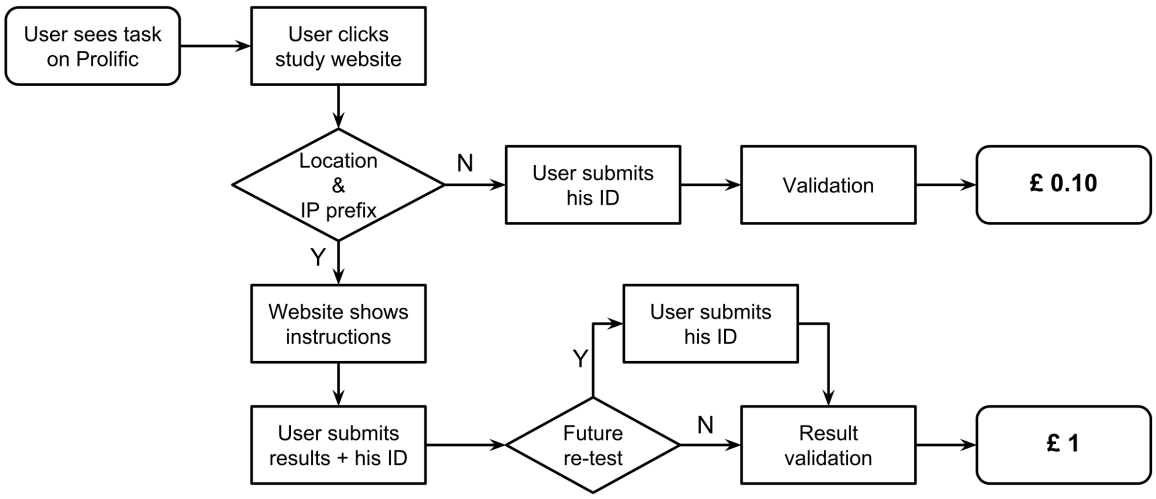
```
> # Variance Inflation Factor
> vif(mydata)
  Variables      VIF
1   SpoofIP 1.651072
2       IDI 3.424585
3       GCI 3.009701
4       GDP 3.154654

AHC <- hclust(dist(mydata, method = "euclidean"), method = "ward.D" )
summary(AHC)

# Plot the obtained dendrogram
plot(hc1, labels = CountryList$Country, ylab = "Eucliden distance", main="", cex
= .6, hang = -1)
rect.hclust(hc1, k = 3, border = 2:4)

# Elbow graph
#install.packages("factoextra")
require(factoextra)
fviz_nbclust(CountryList, hcut, method = "wss") +
geom_vline(xintercept = 3, linetype = 2)
```

Appendix 4: Crowdsourcing measurements

1. Measurement infrastructure

Using crowdsourcing platforms for network measurements presents a series of complications. First and foremost, these platforms are designed to filter users on the base of demographics, not on network properties. Moreover, some platforms do not allow users to download executable for obvious security concerns. Prolific itself allow users to download software, but only after the software has been tested by the staff of the platform. Therefore, we contacted Prolific's staff, who gave us the permission to proceed. We had a several email exchanges with the staff of the platform, who helped us with the various issues that popped up throughout the measurements. In what follows, we briefly describe the set up (also schematise in Figure 27) and the results of the measurements.

Figure 27. Structure of the crowdsourcing measurements.



To facilitate the screening operations and managing users' submissions, we design a website to host the task. This website plays multiple roles. Users from Prolific are redirected on our website, where instruction for downloading and running the test application are shown. Moreover, the website implements the filtering mechanism needed to avoid multiple submissions from the same network, and keeps a record of users viewing out task as well as of the final submissions.

The first important decision refers to the selection of a criteria for filtering users. In fact, without a filtering mechanism, multiple users from the same IP prefix can take the Spoofer test, resulting in a redundancy of measurements and, thus, in a waste of money. The most basic filtering, and the easiest to implement, would be to allow one test per AS. However, for large ASes this type of filtering might

be too restrictive. Therefore, we follow the methodology of our previous experiment: on the base of the size of the AS (measured through the number of IP prefixes announced by the AS), we allow a given number of test. To be more precise, for each AS we compute the size of the IP address space announced, by summing all the possible host IP addresses. Then, we divided the size of the AS by the total IP space for /11 networks (the /11 granularity is based on the previous experiment (Lone et al., in press)). The result of this division gives us the maximum number of /11 prefixes we allow. Therefore, for each AS, we allow a given number of /11 prefixes to be tested.
This operation is automatically done by our website every time a user visits it. Once a user submits the test results, the number of test allowed from that particular prefix is decreased. Naturally, we keep track of users' IP prefix, and relative AS and countries. We do not collect any personal information about users, in accordance with Prolific terms of services. However, to submit the results users are asked to enter their ProlificID, a unique identifier assigned by the platform to users, which is needed for the payment of users.

When a user visits our website, the combination of his IP prefix and AS is checked in our database: if this combination already exists (i.e. that prefix is already tested), or if the maximum number of possible test of that AS is already reached, or if the users comes from a country not included in our country selection, the website display a message explaining that the user is not eligible for our study. Else, the instruction for doing the test and a form to submit test results are presented. To prove that users have indeed taken the test, we require them to submit the URL automatically generated by the Spoofer application pointing to the results of the test on CAIDA's website. Once the user performs the test, we also ask if he would like to take part in the post-test measurements, and, eventually, we record his ProlificID in a separate table.

Finally, we need to decide the reward for users. Prolific requires a minimum compensation of £5 per hour (the platform is based in England and payments are computed in Pounds). During the design of the task, we are asked to estimate the average duration of the task and, on the base of this time, the minimum compensation is computed. We observed that downloading, installing and running the application takes around 5 minutes, which correspond to a minimum reward of £0.42 (€0.48).

Again, we base the structure of the rewards on our previous experience. Since we noticed that not-eligible users tend to complain in the absence of a compensation, we opted for offering not-eligible users a small reward of £0.10, because with a lot of complaints the study might be suspended. For users who are instead eligible, we wanted to create three different pricing strategies: we would start a fist campaign offering £0.50, which we would raise to £0.75 in the second campaign, and, finally, in the third campaign we would offer £1. The logic behind this pricing scheme is that we aim to attract as many users as possible, beginning by all users inclined to do the test for a smaller compensation and incrementally rise the target. Nevertheless, during the studies, we received many complaints because our 5 minutes estimation is too optimistic. Firstly, several users reported that the download of the application might take even more than 5 minutes depending on users' Internet connections. Secondly, users noticed that reading carefully the instruction takes additional time which, added to small errors during the installation, brings the duration of the task to an average of 10 minutes. Moreover, we also noticed that our website was slowed because of to the large volume of traffic. In fact, the screening

mechanism presented above has a high computational cost, especially because for each user it evaluates the size of the AS.
Thus, we had to increase the time allocated and raise the minimum payment, allocating 10 minutes and offering directly £1. Unfortunately, this prevented us from analysing the effect of the pricing elasticity as we planned.

2. Results

We started the crowdsourcing campaigns on January 17th, and we carried our measurements for three weeks, with only exception of the United Kingdom, in which after the initial 5 days we started having only non-eligible participants, point at which the campaign was stopped. We launched new studies every week, in order to constantly push our task at the top of the list of available tasks. In each new study, we blacklisted users that already took part in previous studies, so that to avoid giving multiple bonuses to the same user. Each week, 9 new studies were created (one per each country), each pre-screening users on the base of the target country.

An overview of the results in the 9 countries previously selected is presented in Table 14. The second column, *views*, represents the total number of users that clicked the study on Prolific to see the details. It is interesting to notice that the number of users that viewed the task on Prolific is around half the population claimed by the platform (with the exception of the UK). The third column of TABLE represents the number of non-eligible users that got the £0.10 payment. The column *completed* indicates the number of test collected (/11 prefixes tested). In the following columns, we report the number of ASes tested, the number of new ASes (ASes that were not present in CAIDA's database), the number of new /24 IP prefixes tested and the number of spoofable /24 prefixes identified.

Table 18. Crowdsourcing pre-test results.

Country	Views	Failed	Completed	AS tested	New ASes	New /24	Spoof /24
Canada	284	134	45	25	2	45	3
Australia	96	39	28	12	3	27	3
United Kingdom	683	442	35	16	0	28	2
France	69	29	16	9	1	11	1
Germany	172	55	17	10	0	15	1
Netherlands	95	52	19	11	2	17	0
Turkey	36	18	4	3	0	4	0
Poland	65	37	11	10	4	11	0
Italy	222	207	27	14	3	24	2
Total	1722	1013	202	110	15	182	12

All in all, in this round of pre-test measurements we spent around £350 (roughly €400), to obtain 202 test from 110 different ASes, also including platform fees and (partial) compensation to users who were not able to complete the study in the allocated time or who reported errors in the process (and therefore not recorded on our website). In fact, a small part of users reported problem with the software (multiple download failures, or crash of the Spoofer application itself). These users received £1 if they managed to run the application in a second time, otherwise they received a bonus of £0.50.
Finally, it is unclear why many users, though probably eligible, opted out from the study. It might be that the reward was to loo, or that they did not want to download and run software. Moreover, many users were using mobile devices, for which the Spoofer application is not available. Finally, language barriers might have discouraged some users.

Appendix 5:
Notification to non-compliant operators

Object:
Lack of anti-spoofing filters - vulnerability disclosure

Body:
DON'T LET ATTACKERS SPOOF YOUR IPs!

Every day, attackers exploit the lack of anti-spoofing filtering for their criminal operations. By not preventing IP source address spoofing, you are part of the 8% of network operators in the Netherlands that allows attackers to perpetrate their malicious activities!

In the effort to prevent further attacks based on IP spoofing, cybersecurity researchers from Delft University of Technology have been conducting measurements of compliance with anti-spoofing best practices. In the Netherlands, we have found only 9 Autonomous Systems (8%) that do not correctly implement anti-spoofing filtering, including yours!

You can find a list of the IP prefixes of AS29073 that have shown evidence of IP spoofing in the attachment. More details about the results of our measurements in the Netherlands are available at the webpage https://www.infospoofing.com/nl/ID=38.
On that website, you can also retrieve all the information you need about the best practices to prevent IP spoofing, as well as instructions to be removed from our "spoofing list".

****** PUBLIC GROUP ONLY ******

We made our results publicly available, and we have shared our webpage with CERTs, Network Operators Groups, researchers and bloggers in order to increase the awareness about non-compliant operators.

****

Have you found our notification useful? Or do you have any remark to our type of measurements? We are working to mitigate the consequences of IP spoofing and to make abuse notifications more effective for network operators. Please help us by participating in a 5 minutes anonymous questionnaire at https://goo.gl/forms/th92EZbqQRORp6HM2.

Appendix 6:
Questionnaire to non-compliant operators

Reporting IP spoofing

Please help us to improve cybersecurity research by completing this 5 minutes anonymous questionnaire. All questions are optional, please answer those you feel comfortable with.

Your participation is very important to understand network operators' point of view on IP spoofing and to get feedback on our notification process. Thank you for taking the time to complete this questionnaire.

About IP source address spoofing

- What is the level of awareness about the importance of preventing IP source address spoofing in your organisation?
  - Low: we were not aware/we have rarely heard about the consequences of the lack of anti-spoofing filters
  - Medium: we have often heard about the consequences of the lack of anti-spoofing filters
  - High: we believe that the lack of anti-spoofing filters represents a major problem for Internet security

- What measures has your organisation implemented to prevent IP spoofing in the past?
  - None
  - Other (please describe) _____

- What factors prevented your organisation from correctly implement anti-spoofing filters?
  - We were not aware about the problem
  - Implementing anti-spoofing filters on our network is too complex
  - Implementing anti-spoofing filters on our network is too expensive
  - Implementing anti-spoofing filters brings no advantages to our network
  - Other (please describe) _____

- Is your organisation planning to implement anti-spoofing filters in the future?
  - Yes
  - No
  - Don't know (please describe) _____

- Is there anything you want to tell us about your experience with anti-spoofing filters and about our measurements of compliance?
  - _____

About security notifications

- Is your organisation responsible for managing the IP addresses we reported?
  - No
  - No, but we forwarded the notification to the appropriate entity
  - Yes
  - Yes, but you reached the wrong contact within our organisation

- How valuable do you rate the information provided in our notification?
  - 1 (not valuable) – 5 (very valuable)

- To what extent have you found useful the website displaying the results of our measurements in your country?
  - We did not visit the website
  - We visited the website, but it was NOT useful
  - Somewhat useful
  - Very useful

- Do you think that providing information on the performances of other network operators increases the chances that underperforming operators implement anti-spoofing?
  - Yes
  - No

- Is there anything you want to tell us about our notifications process?
  - _____

Appendix 7: Notification for NOGs

Object:
Let's stop IP spoofing, now!

Body:
Every day, attackers exploit IP spoofing for their criminal operations. Despite being a known vulnerability for at least 25 years, IP source address spoofing still remains a popular attack method for redirection, amplification, and anonymity attacks. This situation persists partially because of the lack of visibility into which operators lack adequate anti-spoofing measures -- that is, their networks are not compliant with BCP38 and related norms.

In the effort to help increase the adoption of anti-spoofing measures, researchers from TU Delft and CAIDA have been conducting measurements on which networks are compliant. We would like to engage the network operator community to reach out to non-compliant operators and instigate remediating actions.

We have created an overview of our findings on networks in the Netherlands. You can see them on our website:
https://www.infospoofing.com/nl/ID=71

The good news is: over 90% of Dutch operators have measures in place against IP spoofing! We would like to ask your help to get the remaining 10% on board. Feel free to share the link to our website or otherwise help mobilize the non-compliant operators. At our site, you can also find more information about anti-spoofing best practices as articulated by the MANRS initiative.

Help us to make the Netherlands a spoofing-free country!

Appendix 8:
Notification for national CERTs

Object:
Vulnerability disclosure - Lack of anti-spoofing filters

Body:
Every day, attackers exploit IP spoofing for their criminal operations. Despite being a known vulnerability for at least 25 years, IP source address spoofing still remains a popular attack method for redirection, amplification, and anonymity attacks. This situation persists partially because of the lack of visibility into which operators lack adequate anti-spoofing measures -- that is, their networks are not compliant with BCP38 and related norms.

In the effort to help increase the adoption of anti-spoofing measures, researchers from TU Delft and CAIDA have been conducting measurements on which networks are compliant. We would like to engage CERTs and the network operator community to reach out to non-compliant operators and instigate remediating actions.

You can find a list of the ASes and IP prefixes from the Netherlands that have shown evidence of IP spoofing in the attachment. We have also created an overview of our findings on networks in the Netherlands. You can see them on our website:
https://www.infospoofing.com/nl/ID=72

The good news is: 90% of Dutch operators have measures in place against IP spoofing! We would like to ask your help to get the remaining 10% on board. Feel free to share the link to our website or otherwise help mobilize the non-compliant operators. At our site, you can also find more information about anti-spoofing best practices as articulated by the MANRS initiative.

Help us to make the Netherlands a spoofing-free country!

Appendix 9:
Notification to security blogs

Object:
Increase awareness about IP spoofing

Body:
Hi,

We are a group of cybersecurity researchers from Delft University of Technology, attempting to incentivize network operators and ISPs to implement BCP38, an important security best practice to prevent attacks based on IP spoofing. IP spoofing has been a well known security issue for more than 25 years, but still, network operators do not implement anti-spoofing filters because of a lack of economic incentives. In fact, doing anti-spoofing is a "good neighbour" policy, which relies on cooperation between operators for their mutual benefit.

We have conducted measurement of compliance with BCP38 to identify which network are compliant and which are not, and we are now trying to “publicly” notify non-compliant operators. We have created an overview of our findings on networks in the Netherlands, you can see them on our website: https://www.infospoofing.com/nl

We are testing whether disclosing this information can be an effective way to notify non-deployer operators, and to get BCP38 implemented. Promoting our website and increasing the visibility of the information we report is a fundamental step for the success of our notifications. Thus, we would like to ask your collaboration to write a post on your blog about it, in order to spread information on non-deployer operators in the Netherlands.

If this might interest you, we would be glad to provide all the information you need. We also have collected some readings on IP spoofing and BCP38 on our website, you can find them here: https://www.infospoofing.com/remediation

Appendix 10:
Logistic analysis on the visits to the website

We model the probability that operators open the link to the website included in the notification as a function of the type of disclosure and on organisational and socio-technical factors:

- *Type of notification*
  - x_1***: Public***
    A binary variable set to 1 for ASes in the public notification group.
- *Organisational factors*
  - x_2***: AS size***
    A continuous variable measured via the number of IP prefixes announced by the AS.
  - AS type
    A categorical variable for the type of the AS, divided in the following binary variables:
    - x_3***: Access provider***
    - x_4***: Enterprise***
    - x_5***: Content provider***
- *Socio-technical factors*
  - x_6***: GDP per capita***
    The GDP per capita of the country of the AS.
  - x_7***: ICT score***
    The score of the country of the AS on the ICT Development Index.
  - x_8***: GCI score***
    The score of the country of the AS on the Global Cybersecurity Index.
  - x_9***: English native country***
    A binary variable set to 1 for ASes in English native speaker countries.

The sample included the 67 ASes in the public and private group.

The results of the logistic regression are presented in TABLE. No predictor has a significant effect on the likelihood that operators visit our website.

To assess the goodness-of-fit of the model, we plot the Receiver Operating Characteristic Curve (ROC), shown in Figure 28. It summarizes the model performance between sensitivity (true positive error rate) and specificity (false positive error rate). Next, we compute the Area Under the Curve, which reveals that the accuracy of our model is poor (69.3% AUC score).

Finally, we compute different pseudo-R^2 parameters, shown in TABLE. All of them present low values, which confirms the poor accuracy of our model.

Table 19. Results of logistic model for visit to the website

	Dependent variable	
	Website Visit	
x_1: Public	0.48	(0.59)
x_2: AS size	-0.40	(0.39)
x_3: Access provider	0.29	(0.98)
x_4: Enterprise	0.78	(1.18)
x_5: Content provider	0.88	(1.14)
x_6: GDP per capita	0.30	(0.64)
x_7: ICT score	0.54	(0.48)
x_8: GCI score	-1.37	(0.88)
x_9: English native country	1.01	(1.22)
Observations	67	
Log-likelihood	-41.1618	

Note: * $p < 0.1$; ** $p < 0.05$: *** $p < 0.01$; Standard error in brackets

Figure 28. Model diagnosis with ROC curve.

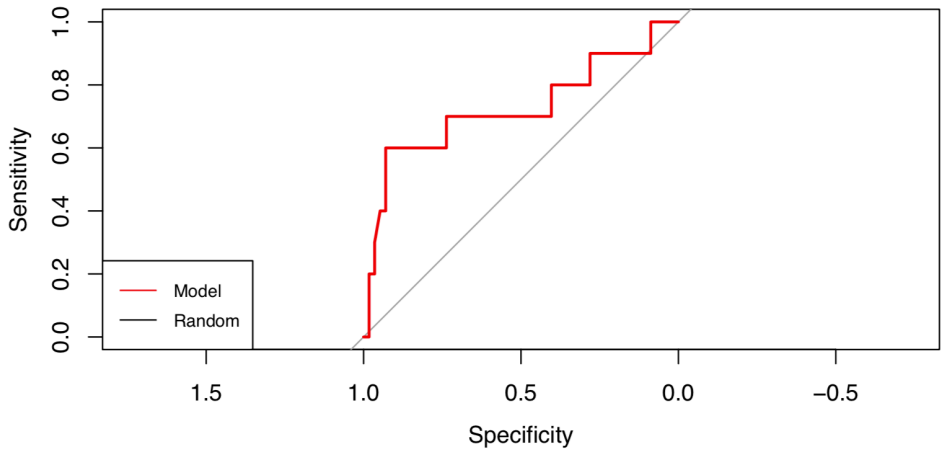


Table 20. Goodness-of-fit of the model

	Pseudo-R^2		
AIC	McFadden	Cox & Snell	Nagelkerke
102.32359	0.088	0.112	0.15

Appendix 11:
R code for logistic analysis

```
> # Data Overview
> summary(data)

      ASN             Visited        Remediation         Private      
 Min.   :   680   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:  9240   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median : 31034   Median :0.000   Median :0.0000   Median :0.0000  
 Mean   : 42662   Mean   :0.403   Mean   :0.1493   Mean   :0.4478  
 3rd Qu.: 45390   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :206360   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  

     Public            IDI             GCI           GDP        
 Min.   :0.0000   Min.   :689.0   Min.   :622   Min.   :12421  
 1st Qu.:0.0000   1st Qu.:824.0   1st Qu.:679   1st Qu.:40341  
 Median :1.0000   Median :839.0   Median :783   Median :42070  
 Mean   :0.5522   Mean   :822.1   Mean   :746   Mean   :40810  
 3rd Qu.:1.0000   3rd Qu.:865.0   3rd Qu.:783   3rd Qu.:45670  
 Max.   :1.0000   Max.   :865.0   Max.   :824   Max.   :49928  

      Size              Content         Enterprise         Access      
 Min.   :       0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:    6912   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :   61440   Median :0.0000   Median :0.0000   Median :1.0000  
 Mean   : 1691745   Mean   :0.1493   Mean   :0.2239   Mean   :0.5224  
 3rd Qu.:  471168   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :34242560   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  

   English      
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :1.0000  
 Mean   :0.5373  
 3rd Qu.:1.0000  
 Max.   :1.0000  

> head(data)
   ASN Visited Remediation Private Public IDI GCI   GDP     Size Content
1  680       1           0       1      0 839 679 42070  8307968       0
2 1221       0           0       1      0 824 824 49928 13746028       0
3 1267       0           0       0      1 704 626 30675  6110720       0
4 2856       0           0       0      1 865 783 40341 11216640       0
5 3320       0           0       1      0 839 679 42070 34242560       0
6 4739       0           0       1      0 824 824 49928  1979136       0

 Enterprise Access English
1          1      0       0
2          1      0       1
3          1      0       0
4          1      0       1
5          1      0       0
6          1      0       1
```

```
> # Scale variables
> size <- scale(Size)
> gdp <- scale(GDP)
> idi <- scale(IDI)
> gci <- scale(GCI)
> access <- scale(Access)
> enterprise <- scale(Enterprise)
> content <- scale(Content)
> english <- scale(English)
> visited <- scale(Visited)

> # Logit model
> logit <- glm(Remediation ~ size + access + enterprise + content + gdp + idi +
gci + english + visited, family=binomial (link="logit"))

> summary(logit)

Call:
glm(formula = Remediation ~ size + access + enterprise + content +
    gdp + idi + gci + english + visited, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.60149  -0.14568  -0.04961  -0.00530   1.90996

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.83723  399.23343  -0.015   0.9883
size          0.25295    1.08182   0.234   0.8151
access        7.61482 1923.09392   0.004   0.9968
enterprise    6.87178 1604.86604   0.004   0.9966
content       7.60451 1371.92040   0.006   0.9956
gdp           0.05905    1.23496   0.048   0.9619
idi          -1.73938    1.45358  -1.197   0.2315
gci           3.11566    2.07859   1.499   0.1339
english      -2.95055    1.75712  -1.679   0.0931 .
visited       2.32581    0.99522   2.337   0.0194 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 56.469  on 66  degrees of freedom
Residual deviance: 22.840  on 57  degrees of freedom
AIC: 42.84

Number of Fisher Scoring iterations: 18

> # Confidence Intervals
> confint(logit)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept)  -788.3203788  776.6459179
size           -1.8673839    2.3732829
access      -3761.5800113 3776.8096449
enterprise  -3138.6078714 3152.3514221
content     -2681.3100686 2696.5190868
gdp            -2.3614213    2.4795247
idi            -4.5883499    1.1095821
gci            -0.9582924    7.1896117
english        -6.3944316    0.4933363
visited         0.3752186    4.2763982
```

```
> # Log likelihood
> logLik(logit)
'log Lik.' -11.41977 (df=10)

> # Odd Rations
> exp(coef(logit))
 (Intercept)         size       access   enterprise      content          gdp
2.916910e-03 1.287818e+00 2.028023e+03 9.646597e+02 2.007226e+03 1.060830e+00
         idi          gci      english      visited
1.756286e-01 2.254830e+01 5.231105e-02 1.023495e+01

> # Predicted probabilities
> plogit <- predict(logit, type="response")
> summary(plogit)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000000 0.0003918 0.0059847 0.1492537 0.1609331 0.9982316
> table(true = Remediation, pred = round(fitted(logit)))
    pred
true  0  1
   0 54  3
   1  4  6

># ROC curve
> ROC <- roc(Remediation ~ plogit, data = data)
> plot(g, col=c("red2", "black"))
> legend("bottomleft", legend=c("Model", "Random"), col=c("red", "black"),
lty=c(1,1), cex=0.79)

> AUC
> auc(ROC)
Area under the curve: 0.9588
> PseudoR2(logit)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
       0.5955404        0.2059479        0.3946426        0.6929544
McKelvey.Zavoina           Effron            Count        Adj.Count
       0.9178086        0.5369656        0.8955224        0.3000000
             AIC    Corrected.AIC
      42.8395375       46.7681089
```