Increasing the Impact of Voluntary Action Against Cybercrime

Çetin, Orçun

DOI
10.4233/uuid:ad5d9147-b3ef-4708-b954-142b00820499

Publication date
2020

Document Version
Final published version

Citation (APA)
Çetin, O. (2020). *Increasing the Impact of Voluntary Action Against Cybercrime*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ad5d9147-b3ef-4708-b954-142b00820499

Increasing the Impact of Voluntary Action Against Cybercrime

DISSERTATION

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Wednesday 22 January 2020 at 12:30 o'clock

by

Feyzullah Orçun ÇETIN

Master of Science in Networks and Security, University of Kent,
United Kingdom
born in Izmir / Turkey

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr. M.J.G van Eeten	Delft University of Technology, promotor
Dr.ir. C. Hernandez Ganan	Delft University of Technology, copromotor

Independent members:

Prof.dr.ir. H.J. Bos	Vrije Universiteit Amsterdam
Prof.dr. P.H. Hartel	Delft University of Technology
Prof.dr.ir. J. Hernandez-Castro	University of Kent, United Kingdom
Prof.dr.ir. P.H.A.J.M. van Gelder	Delft University of Technology
Prof. dr. W. P. Stol	Open University of the Netherlands

This research has been funded by Netherlands Organisation for Scientific Research (NWO) (grant nr. 12.003/628.001.022).

Distributed by Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628BX Delft, the Netherlands.

Keywords: cybersecurity, network security, security economics, hosting provider, ISP, incentives, domain owners, patching, vulnerability scan, abuse notifications, vulnerability notifications walled garden, data analysis, statistical models.

Summary

Resources on the Internet allow constant communication and data sharing between Internet users. While these resources keep vital information flowing, cybercriminals can easily compromise and abuse them, using them as a platform for fraud and misuse. Every day, we observe millions of internet-connected resources are being abused in criminal activities, ranging from poorly-configured Internet of Things (IoT) devices recruited into flooding legitimate services' networks with unwanted Internet traffic or compromising legitimate websites to distribute malicious software that is designed to prevent access to victim's data or device until a ransom has been paid to the attacker.

The Internet's decentralized architecture necessitates that defenders must voluntarily collaborate to combat cybercrime. While mandatory efforts may be necessary in some circumstances, the bulk of incident response will remain based on voluntary actions among thousands of Internet intermediaries, researchers and resource owners. These voluntary actions typically take the form of one party sending security notifications to another about potential security issues and asking them to act against it. Security notifications are intended to support and promote a wide range of feasible efforts, which aim to detect and mitigate millions of daily incidents and remediate underlying conditions. Despite its importance, voluntary action remains a poorly understood and significantly less investigated component of the fight against cybercrime. All of this puts a premium on understanding how voluntary cyber-defense efforts prove to be the most effective in remediating security issues. Thus, this leads to the main research question of the thesis:

How can the effectiveness of voluntary action against cybercrime be increased?

This research question required us to systematically analyze the relationship between characteristics of notification mechanisms and security issues at the key Internet intermediaries, such as Internet service providers and hosting providers. We investigated this relationship by measuring remediation rates of security issues after sending security notification. All of the studies have been well received by both academia and the industry. Some of their findings have become starting points for the next research step towards a more secure Internet. The research starts with measuring a hosting provider's ability to remediate compromised websites in their network. These websites were compromised and abused by the attackers to be used as phishing websites. We know remarkably little about the factors that drive higher response rates to abuse reports. One such factor is the reputation of the sender.

In Chapter 2, we present a study that measures the impact of abuse notifications and a notification sender's reputation on compromised cleanup rates. In the first part of the

study, we measured the effectiveness of the abuse notifications by comparing two groups of compromised websites. One group received abuse notifications, and the other did not. In the second part of the study, we assess the effectiveness of issuing notifications from three senders with different reputations: an individual, a university and an established anti-malware organization. Additionally, we also studied the efficacy of cleanup advice provided via a link in the notifications. Our results showed that abuse reports significantly increase the remediation rates compared to not notifying. However, sender reputation did not significantly influence the cleanup process. Furthermore, our results suggest that providing a cleanup website containing specific instructions improves the cleanup speed when hosting providers view the instructions.

In Chapter 3, we investigated intermediaries' and resources owners' ability to remediate vulnerabilities. Our study investigated the effectiveness of reaching out to different affected parties, and once reached incentivize for vulnerability remediation. The study compared the effectiveness of direct and intermediary remediation strategies in terms of remediation and reachability to find out which channel mobilizes the strongest incentive for remediation. Results demonstrated that there is no good communication mechanism for getting the wealth of vulnerability remediation information to the affected parties. Additionally, we studied whether providing a link to a mechanism to verify the existence of the vulnerability could incentivize resource owners and intermediaries to act upon our notifications. Our results showed no evidence that notifications with vulnerability demonstrations did better than standard notification for both resource owners and intermediaries.

After investigating the effectiveness of notifications made to vulnerable and compromised websites owners and intermediaries, we collaborated with an ISP to measure the effectiveness of notifications made to vulnerable and infected device owners. In Chapter 4, we studied user behavior and remediation effectiveness of an alternative mechanism for notification and remediation: quarantining the resource in a so-called walled garden environment. We studied the relationship between cleanup rates and other factors, such as the release mechanism used to get out of quarantine, and the time spent in a quarantine environment. Our results illustrate that almost three-quarters of the quarantined users had managed to clean their infected machines in their first two attempts of quarantining when they have an option to self-release themselves from the quarantine environment. Significantly, providing an option to self-release from the quarantine environment did not introduce lax security behavior.

In Chapter 5, we assess the effectiveness of the walled garden by comparing remediation with two other groups: one group which was notified via email but not quarantined and another group where no action was taken. Our results found very high remediation rates for the quarantined users, even though they can self-release from the quarantine environment. Moreover, the walled garden group achieved higher remediation rates than both email and control groups. Surprisingly, over half of the customers who were not notified at all also remediated, though this is tied to the fact that many observations of vulnerable servers are transient.

With the rise of IoT malware, cleaning up infected devices in ISP networks has become a critical task. In Chapter 6, we presented remediation rates from an observational study and a randomized controlled trial involving 220 consumers who suffered from Mirai infection. Our findings showed that walled garden notifications achieved higher Mirai malware remediation rates than email notifications. Moreover, our results showed that email notifications have no observable impact compared to a control group where no notifications were sent. However, improving the content of the walled garden notification with more actionable content did not increase the remediation rates.

Our research provides a better understanding of how effective these actors are in terms of abuse and vulnerability remediation and how can they be more effective in hosting and ISP market. Concerning the implications of our results for practice, I conclude that voluntary action can be improved by understanding and improving the incentives of Internet intermediaries and resource owners. Both laws and softer governmental mechanisms can be used to incentivized resource owners and intermediaries to act more effectively against cybercrime.

Samenvatting

Computers op internet maken communicatie en gegevensuitwisseling tussen internetgebruikers mogelijk. Hoewel deze computers vitale informatiestromen ondersteunen, kunnen cybercriminelen ze compromitteren en ze gebruiken als een platform voor fraude en misbruik. Elke dag zien we dat miljoenen met internet verbonden computers worden misbruikt bij criminele activiteiten, variërend van slecht geconfigureerde Internet of Things (IoT)-apparaten waarmee grote hoeveelheden ongewenst internetverkeer worden afgevuurd op doelwitten tot het overnemen van legitieme websites om schadelijke software te verspreiden, zoals ransomware of spyware.

De gedecentraliseerde architectuur van internet, over talloze landsgrenzen heen, vereist dat verdedigers vrijwillig moeten samenwerken om cybercriminaliteit te bestrijden. Hoewel in bepaalde omstandigheden verplichte inspanningen opgelegd kunnen worden, blijft het grootste deel van de respons op incidenten gebaseerd op vrijwillige acties van duizenden internetbemiddelaars, onderzoekers en eigenaars van hulpcomputers. Deze vrijwillige acties nemen meestal de vorm aan van een partij die beveiligingsproblemen detecteert en aan een andere partij meldt, met de vraag om hiertegen op te treden. Deze meldingen worden 'abuse reports' genoemd. Dagelijks worden miljoenen abuse reports verstuurd om op die manier de gedetecteerde incidenten verholpen te krijgen.

Ondanks het belang ervan, blijft vrijwillige actie een slecht begrepen en weinig onderzocht onderdeel van de strijd tegen cybercriminaliteit. Dit alles draagt ertoe bij dat we willen begrijpen hoe vrijwillige inspanningen op het gebied van cyberverdediging effectiever gemaakt kunnen worden bij het oplossen van beveiligingsproblemen. Dit leidt tot de centrale onderzoeksvraag van het proefschrift:

Hoe kan de effectiviteit van vrijwillige actie tegen cybercriminaliteit worden verhoogd?

Voor deze onderzoeksvraag moesten we de relatie tussen kenmerken van de meldingsmechanismen en beveiligingsproblemen bij de belangrijkste internet-intermediairen, zoals internetproviders en hostingproviders, analyseren. We hebben deze relatie onderzocht door het herstelpercentage van beveiligingsproblemen te meten na het verzenden van een beveiligingsmelding (abuse report). Deze uitkomsten hebben aandacht gekregen van zowel de wetenschap als van de industrie. Sommige van hun bevindingen zijn uitgangspunten geworden voor verbeteringen in de meldingsmechanismen voor een veiliger internet.

De eerste studie meet de mate waarin hostingprovider gecompromitteerde websites in hun netwerk herstellen als ze hierover een melding hebben ontvangen. Deze websites zijn

door de aanvallers gecompromitteerd en misbruikt om als phishing-websites te worden gebruikt. We weten opmerkelijk weinig over de factoren die een hogere respons op misbruikmeldingen veroorzaken. Een van die factoren is de reputatie van de afzender. In hoofdstuk 2 presenteren we een studie die de impact meet van meldingen en van de reputatie van een afzender op opschoonpercentages. In het eerste deel van het onderzoek hebben we de effectiviteit van de misbruikmeldingen gemeten door twee groepen gecompromitteerde websites te vergelijken. De ene groep ontving misbruikmeldingen en de andere niet (deze functioneerde als controlegroep). In het tweede deel van het onderzoek vergelijken we de effectiviteit van meldingen van drie afzenders met verschillende reputaties: een individu, een universiteit en een bekende anti-malware-organisatie. Daarnaast hebben we ook de effectiviteit bestudeerd van de opschoonadviezen die beschikbaar waren gesteld via een link in de melding. Onze resultaten toonden aan dat misbruikrapporten de saneringspercentages aanzienlijk verhogen in vergelijking met het niet melden. De reputatie van de afzender had echter geen significante invloed op het opruimproces. Bovendien suggereren onze resultaten dat het aanbieden van een opschoonwebsite met specifieke instructies de opschoonsnelheid verbetert wanneer hostingproviders de instructies bekijken.

In hoofdstuk 3 onderzochten we of meldingen aan intermediairs (hostingaanbieders en netwerkbeheerders) helpen om kwetsbaarheden in computersystemen te verhelpen. We verzamelden data over onbekende kwetsbaarheden in DNS servers. Vervolgens onderzochten we of we de getroffen partijen konden bereiken en, als deze werden bereikt, of ze de kwetsbaarheid ook daadwerkelijk verhielpen. We vergeleken de effectiviteit van meldingen direct aan de (vermeende) eigenaar van de DNS server met een melding aan een intermediair, om erachter te komen welk kanaal de sterkste prikkel voor herstel mobiliseert. De resultaten toonden aan dat er geen goed communicatiemechanisme bestaat om de rijke informatie aangetroffen kwetsbaarheden bij de betroffen partijen te krijgen. Daarnaast hebben we onderzocht of het aanbieden van een link naar een site die demonstreert dat het beveiligingslek daadwerkelijk aanwezig is in de betreffende server helpt om eigenaren tot actie te bewegen. Onze resultaten toonden geen bewijs dat meldingen met kwetsbaarheidsdemonstraties het beter deden dan standaardmeldingen voor zowel de eigenaren van de server als voor intermediairs.

Voor de volgende studie hebben we samengewerkt met een Internet Service Provider (ISP), oftewel internetaanbieder. We wilden de effectiviteit meten van meldingen van de ISP aan de consumenten met kwetsbare of geïnfecteerde apparaten. In hoofdstuk 4 hebben we de effectiviteit een bijzonder meldingsmechanisme bestudeerd: de verbinding van de consument met een besmet apparaat wordt in quarantaine geplaatst (in een zogenaamde 'walled garden'). We onderzochten de relatie tussen opschoonpercentages en andere factoren, zoals het mechanisme dat de getroffen consument kan gebruiken om uit de quarantaine te komen en de tijd die wordt doorgebracht in een quarantaineomgeving. Onze resultaten tonen aan dat bijna driekwart van alle in quarantaine geplaatste gebruikers erin is geslaagd hun geïnfecteerde machines op te schonen in hun eerste twee pogingen, wanneer ze een optie hebben om zichzelf uit de quarantaineomgeving te bevrijden. Veelzeggend is

dat het bieden van een optie voor eigenhandige vrijgave uit de quarantaineomgeving geen laks beveiligingsgedrag met zich meebracht.

In hoofdstuk 5 beoordelen we de effectiviteit van het quarantaine-mechanisme door twee andere groepen te vergelijken: een groep die via e-mail op de hoogte is gesteld versus een groep die in quarantaine is geplaatst (en een controlegroep waar niet meteen actie is ondernomen). Onze resultaten toonden zeer hoge herstelpercentages voor in quarantaine geplaatste gebruikers, ook al konden zij zichzelf vrijgeven uit de quarantaineomgeving. Dit was hoger dan zowel de e-mailgroep als de controlegroep. Verrassend genoeg is meer dan de helft van de klanten die niet op de hoogte ook van de besmetting bevrijd. Dit kan mogelijk verband houden met het feit dat veel kwetsbare systemen slechts tijdelijk besmet zijn.

Met de opkomst van IoT (internet of things) apparaten is ook IoT malware opgekomen, zoals Mirai. Daarmee wordt ook het opruimen van geïnfecteerde IoT apparaten in ISP-netwerken een cruciale taak. In hoofdstuk 6 presenteren we de herstelpercentages van een studie en experiment met 220 consumenten bij wie sprake was van een Mirai-infectie. Onze bevindingen toonden aan dat het quarantaine-mechanisme hogere opschoonpercentages behaalden dan e-mailmeldingen. Bovendien hebben onze resultaten aangetoond dat e-mailmeldingen geen hogere effectiviteit halen dan de controlegroep die pas later de meldingen heeft ontvangen.

Ons onderzoek geeft beter inzicht in hoe effectief actoren in de hosting en ISP-markt zijn in het herstellen van kwetsbaarheden en besmettingen. Wat betreft de implicaties van onze resultaten voor de praktijk, concluderen wij dat vrijwillige actie kan worden verbeterd door de prikkels van intermediairs en eigenaren van computers te versterken. Het proefschrift concludeerde dat de overheid zowel wetten kan inzetten als 'zachtere' mechanismen om eigenaars van computers en intermediairs als hostingbedrijven en ISP's te stimuleren om effectiever op te treden tegen cybercriminaliteit.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Problem statement

Resources on the Internet allow constant communication and data sharing between Internet users. While these resources keep vital information flowing, cybercriminals can easily compromise and abuse them, using them as a platform for fraud and misuse. There are various means to misuse an Internet-connected resource, some more damaging than others. Among these are compromising a resource to steal credit card information, making unauthorized purchases or attacking others by forcing the resource to send unwanted Internet traffic.

In 2017, a compromised network of Internet-of-Things (IoT) devices, ranging from home routers to security cameras, almost brought down the Internet by launching a series of powerful distributed denial of service (DDoS) attacks, in which targets were simply flooded with web traffic until they were swamped and knocked offline [1]. Some of the targeted companies reported attack volumes significantly higher than what was observed from previous attacks. These attacks were carried out by malicious software, commonly known as malware. This particular piece of malware was called Mirai, along with its variants [2]. Compromised devices carry on compromising other devices by simply guessing their login credentials, which are usually factory default usernames and passwords[3]. Once the password is guessed, a malicious file is inserted which takes control of the device to use it for malicious purposes. One variant of Mirai caused a significant outage for one of the largest German Internet Service Providers (ISPs) while looking for insecure devices to compromise [4]. In this attempt to compromise vulnerable routers, more than 900,000 customers were affected [5]. The same attack also knocked thousands of Internet users offline in other ISP networks.

Similarly, vulnerable web servers are often targeted by attackers to deliver malicious software or fraudulent pages that trick visitors into sharing their sensitive information. In late 2018, attackers compromised thousands of websites running vulnerable and outdated Wordpress themes and plugins [6]. Malicious code was inserted into the pages of the compromised websites which then redirected the visitors of the compromised websites to fraudulent sites claiming to be Microsoft technical support. Owners of affected websites

and hosting providers had to clean up the malicious code and address underlying issues that caused the website to be vulnerable. As the incidents above demonstrate, millions of vulnerable resources, ranging from IoT devices to web servers and computers connected to the Internet are being regularly compromised and abused by cybercriminals to be used as a platform to attack others or for financial gain [7, 8, 9, 10].

A safer Internet ecosystem requires continual detection and remediation of compromised and vulnerable resources. This process consists of 4 significant steps: detection of the security problem, identifying the solution for the security problem, identifying affected parties and lastly notifying affected parties to start the remediation process. Research on detecting cybersecurity problems and finding remedies has advanced significantly, with security researchers discovering and patching thousands of new vulnerabilities each year [11]. Similarly, thousands of unique malicious software indicators are discovered and blocked every day [12]. Furthermore, large scale discovery of thousands of malicious and vulnerable resources has become fairly straightforward with new scanning tools and techniques [13, 14, 15, 16]. However, these have a very limited impact on our ability to determine effective ways to notify and provide incentives to those who can remediate vulnerable and abused resources. Thus, the majority of the resources remain vulnerable or compromised months after the discovery of the security issue and their solutions [17, 18].

A variety of actions can be taken to deal with vulnerable and abusive hosts on the Internet. Some of these actions are mandatory and enforced by governmental agencies, while others are voluntary. Although formal and mandatory actions are essential to fight against cybercrime, the bulk of these actions are voluntary actions of many thousands of private actors. Typically, these actors are ranging from researchers and security companies that are willing to share incident and vulnerability data with relevant Internet intermediaries that facilitate the use of the Internet and subscribers of these resources and services. Voluntary collaboration among these actors is crucial in cleaning up malware-infected resources and preventing them from being easy targets for criminals. For example, the DNS Changer Working Group notified various ISPs to clean up a group of computers that had been infected by a malware family and had come under the control of malicious actors [19]. Similarly, the Conficker Working Group coordinated with registrars to shut down domain names used to control another group of malware-infected machines [20]. In both of these cases, working groups and partnering Internet intermediaries voluntarily committed to remediating malicious resources. Despite its importance, voluntary action remains a poorly understood and significantly less investigated component of the fight against cybercrime.

1.2 Abuse and Vulnerability Reporting[1]

To better understand how security problems are reported and remediated, we developed a framework model illustrated in Figure 1.1. This model is an improved version of the earlier

[1]Part of this section is based on previous work on abuse reporting infrastructure [21].

effort [21] where abuse reporting and remediation are described. In this model, we describe how abuse and vulnerabilities are remediated or used for protecting resource owners. Descriptions and examples of both actors and actions are provided in Table 1.1. The model displays three key components of abuse and vulnerability reporting infrastructure: abuse and vulnerability data collection, dissemination and lastly remediation and protection.

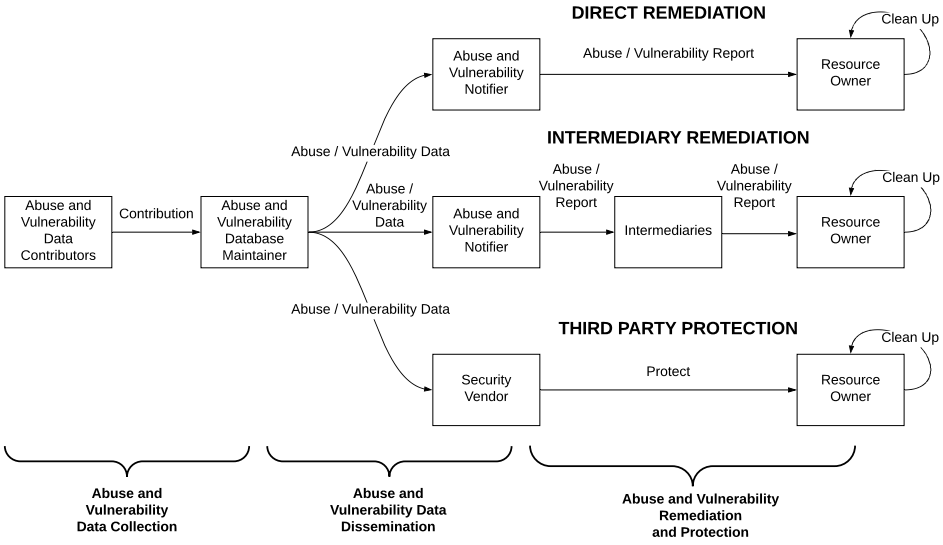


Figure 1.1: Abuse and vulnerability reporting infrastructure overview

Abuse and vulnerability data collection

Remediation and protection are the two main reasons why abuse and vulnerability data is collected. The purpose of the remediation is to eliminate security problems cornering the online resources by reaching the resource owner or intermediary. This security concern can be a vulnerability that could lead to a compromise or malicious behavior that causes harm to third parties or the users of the compromised resource. Abuse notifications are sent to deal with malicious and abusive behavior. This can be done by removing the malicious code and files placed or suspending the resource's network access. These notification efforts treat the symptoms of the underlying problem by disabling malicious behavior rather than going after the root causes. To deal with the root causes of the problem, vulnerability reports are disseminated. Vulnerability reports target the vulnerabilities and misconfigurations that

Table 1.1: Actors and actions in the reporting infrastructure model

Actors	**Role**	**Example**
Abuse and Vulnerability Data Contributor	Identifies and reports instance of abusive and vulnerable host	End users who is reporting suspicious emails in their inbox
Abuse and Vulnerability Database Maintainer	Receives and aggregates abuse and vulnerabilities data	PhishTank [22]
Abuse and Vulnerability Notifier	Entity to disseminate collected abuse and vulnerability data	Google Safe Browsing [23] Shadowserver [24]
Intermediary	Facilitate the use of the Internet and remediation for their subscribers	ISPs, Hosting providers
Resource Owner	Owners of the resource, responsible for the remediation	IoT device owners, Domain owners
Security Vendor	Uses abuse and vulnerability data to protect their subscribers	Mozilla Firefox[25], McAfee [26]
Actions	**Description**	**Example**
Contribute	Proof of abuse or vulnerability that are provided to abuse and vulnerability database maintainer	Submitting suspicious emails to Google [27]
Send Abuse and Vulnerability Data	Transforming raw data into actionable intelligence	Sending list of compromised websites to Shadowserver
Send Security Report	Sending abuse and vulnerability reports to intermediaries, resource owners and security vendors	Sending notifications to ISPs when one of its customers is infected
Protect	Using abuse and vulnerability data to protect subscribers	Blocking list of compromised websites

cause the abuse in the first place. Typically, affected parties can tackle the vulnerabilities by patching the resource or making the vulnerable resources inaccessible to the abusers and third parties. On the contrary, protection does not concern itself with resolving the security issues. The main purpose of protection is to defend the resources of third parties against harm caused by abuse or malicious resources on the Internet rather than remediating them. This promotes a strong incentive to collect abuse and vulnerability data because it can be sold as a service to third parties.

In practice, abuse and vulnerability data is collected either manually or automatically. In the manual collection, abuse and vulnerability reports are manually entered and forwarded to abuse and vulnerability data maintainers. Data collectors can be security professionals from security companies or even Internet users. For example, an Internet user might notice a suspicious email asking them to provide their bank account credentials. The user then manually submits this information to an abuse data maintainer such as PhishTank [22]. Sometimes, financial institutions discover web pages that impersonate their websites and products from individual user reports. Internet users do not need to be working for these organizations or even using their products and services in order to make a report. Many organizations offer pages to receive manually generated security reports from Inter-

net users. Victimized organizations can share this collected abuse data with abuse and vulnerability database maintainers. Often manual contributions are based on a small number of contributions. Even when security professionals and volunteers are involved their reporting ability does not scale much.

To increase the quantity of abuse and vulnerability data contributions, security companies might deploy automated tools for abuse and vulnerability discovery. For instance, automated report generation can be achieved by a vulnerable system set up as a decoy to lure attackers. Attempts to gain unauthorized access can be used to generate reports for abuse and vulnerability database maintainers. Similarly, automated tools can be programmed to actively look for security problems rather than passively waiting for them to occur. Vulnerabilities and various abusive hosts are often collected in an automated manner. There are several methods and tools to find vulnerabilities and abusive behavior on the Internet. Consider the example of Google's search engine security scans. These scans can detect and report misbehaving websites automatically. As a result of these scans, misbehaving websites can be easily found.

Abuse and vulnerability data dissemination

Dissemination begins when abusive and vulnerable hosts are detected and sent to a database maintainer. Typically, the database maintainer maintains an open channel to receive abuse and vulnerability data from contributors, which can be in the form of raw indicators such as suspicious emails. Database maintainers might further process the contributed data to produce a more actionable data set. For example, spam emails generally contain links to compromised websites. When a suspicious email is contributed, links to compromised websites need to be extracted carefully. After further processing, abuse and vulnerability data is aggregated into blacklists or more comprehensive reports. Subsequently, the maintainer's duty is to deliver this data into the hands of notifiers to promote remediation or protection against potentially malicious Internet resources.

Generally, distinct entities are carrying out roles of maintainer, protector or notifiers. For instance, the PhishTank initiative can be given as an example of distinct abuse data maintainers. PhishTank allows organizations and individuals to submit links of websites that are observed in imitating well-known company or organization websites. PhishTank data is used by notifiers to remediate the abusive resources by notifying resource owners and intermediaries. Security vendors can use this data in a similar manner to protect their clients. Moreover, it is also possible that all these roles are handled by a single entity. One of the most well-known examples of this is Google Safe Browsing. This initiative operates as both an abuse data maintainer and notifier. As a data maintainer, Safe Browsing can be queried to block misused websites that distribute malware or impersonated websites. As a notifier, the Safe Browsing initiative provides information to hosting providers and ISPs by sending email alerts to system operators regarding abusive resources hosted on their networks [23]. Another well-known example is Shadowserver, which is a non-profit security

organization that gathers and disseminates abuse and vulnerability data to ISPs, hosting providers and other types of network operators [24]. Their reports allow intermediaries to monitor and remediate security problems in their network.

Abuse and vulnerability remediation and protection

Figure 1.1 presents 3 intervention strategies for reported vulnerable and abusive hosts: direct remediation, intermediary remediation, and third party protection.

Direct remediation occurs when the owner of the resource is notified directly by a notifier in the hope that the owner resolves the security problem. Direct remediation requires a channel that can be used to notify resource owners. Typically, notifiers who foster direct remediation are the services that are used by the resource owners themselves. These services have up-to-date contact information of the resource owner to send email notifications. For example, Google's Search Console subscribers receive security notifications directly to their email accounts when a security issue is detected on their websites by Google.

In many cases, direct remediation strategies can be impractical to follow. The resource owners can be impossible to reach or lack the necessary expertise to remediate the security problem at hand [28]. In these cases, notifiers send notifications to intermediaries that give Internet access to online resources. Asking intermediaries to act promotes intermediary remediation. This is an important strategy because intermediaries remediate security problems themselves or forward the notifications to the resource owners in the hopes that it would trigger a remediation action. For example, a hosting provider can be notified by a security company when one of its customers is compromised. Similarly, the same channel is used for notifying vulnerabilities in the hosting provider's network.

On the other hand, third-party protection occurs when vulnerability and abuse data are used by the security vendors to protect their subscribers. Typically, security vendors protect their subscribers by blocking potentially harmful resources collected by the abuse and vulnerability database maintainer. For example, many security companies block Internet traffic originating from compromised and vulnerable devices because these devices can be used as a platform to attack online resources.

While third-party protection does not directly facilitate cleanup, it provides a strong incentive to collect both abuse and vulnerability data because security companies can sell protection as a service. As a result of this, we included protection in the framework. However, aspects and effectiveness of protection mechanisms are not within the scope of this thesis.

1.3 Voluntary action

A secure Internet ecosystem relies to a large extent on security notifications and voluntary action supported by Internet intermediaries and resource owners which have direct or in-

direct access to resources often targeted by the attackers. These voluntary actions typically take the form of one party sending security notifications to another about potential abuse or vulnerability and asking them to act against it.

Security notifications are intended to support and vitalize a wide range of feasible efforts to detect and mitigate incidents and remediate underlying conditions. Every day, millions of security notifications are sent and forwarded to intermediaries (such as hosting providers or ISPs) and resource owners (such as website owners or admins) in the hope that they would act upon it [29, 30, 31]. These intermediaries regularly process security notifications and assist one or more of their customers that are facilitating abuse or vulnerable software. For instance, a mid-sized hosting provider can easily receive hundreds of abuse complaints and forward these complaints to their subscribers to facilitate cleanup each day. In another example, proactive ISPs assist their subscribers to maintain the security of their home devices by voluntarily forwarding the notifications they receive from notifiers. In some cases, intermediaries can voluntarily clean up the resource themselves or temporarily make the resource unavailable until it gets fixed by the resource owners.

Generally speaking, security notifications that drive voluntary cyber-defense efforts can be transmitted using the following methods: pull or push. Proactive service providers and some resource owners tend to pull ongoing updates as a result of security incidents and vulnerability reports as they are detected in their resources. For example, hosting providers may subscribe to a blacklist provider that collects IP addresses used in malicious activities. When utilized by the hosting providers, the blacklist provider provides all malicious IP addresses that belong to their leased range to trigger a cleanup. In the majority of cases, notifications are pushed by the notifiers to affected parties. Email is the most commonly used method to push security notifications because it is cost-effective and scales reasonably well. Typically, publicly available abuse contacts are used to reach the affected parties via email notifications.

In an ideal world, intermediaries and resource owners would act upon all the notifications they receive and subscribe to clearinghouses to identify vulnerabilities and malicious activities in their network to remediate it. Additionally, they should be able to detect vulnerabilities and any kind of misuse in their network so they can perform various actions to mitigate it. However, in practice, many security notifications do not even reach the affected parties due to spam filters, mismanaged email accounts or absence of contact information for the responsible party. Moreover, even when a notification reaches the affected parties, it might not trigger any action. This might be because notifications that were received by the intermediaries and resource owners are simply ignored, overlooked or might not be actionable. Furthermore, in many cases, contacting resource owners would be ineffective. This is because they might lack the technical expertise to remediate the vulnerability or act against the abuse. Additionally, abusive resources might be registered by attackers to be used for malicious purposes. In these cases, notifying an intermediary is a far better option since the intermediary could reach the resource owner by using private information. In the case of there being no reaction to the notifications, the intermediary could simply stop the online

presence of the misbehaving resource. On the other hand, the response of the intermediaries is heavily influenced by their type and business model. Some intermediaries receive security notifications and choose not to react due to the associated extra cost of notifying the customers and the higher cost of network security equipment. Additionally, there is no central authority that verifies the validity of the security notifications. As a result of this, there is no way to verify the validity of the content of the notifications without investigating it. Nonetheless, thousands of security notifications are sent across the Internet without having an established relationship.

There is no legal course of action to persuade, nor legal authority to complain when a security notification is ignored. However, many security reports are acted upon without any strong legal obligation, across various jurisdictions without any pre-established relationship between the notification sender and the recipients. Typically, proactive providers and voluntary initiatives mobilize the whole market in better dealing with security problems. This shows that many companies are making an effort that they are not legally required to do. All of this puts importance on understanding the myriad ways cybercrime notifications are used to identify why defenses do or do not work, and how they might be improved.

1.4 Security incentives of intermediaries and resource owners

Technical advancements alone have proven inadequate in the fight against cybercrime. This is because the extent of action against cybercrime is heavily determined by the incentives of the intermediaries and resource owners. Thus, attempts to remediate issues related to cybercrime also have to take into account the incentives of the key actors that are involved. As a result of this, understanding issues of misaligned incentives among key actors is as significant as improving the technology addressing cybercrime. There are many factors that play a major role while an affected party deals with the vulnerability or the abuse. Most notably, the abuse and vulnerabilities generally do not directly harm the intermediaries or resource owners. For instance, when a web server is hacked by an attacker to be used as a phishing site, the hosting provider is not affected directly. As the examples above demonstrate, when harm associated through cybercrime is indirect, the incentive to fight against cybercrime becomes weaker. Generally, intermediaries avoid harm when there are negative externalities from a lack of security or human error.

Similarly, harm might not be visible to the resource owner or intermediaries. For instance, attackers can upload their malicious pages, separate from legitimate pages, to serve as a phishing platform for victims that were tricked through phishing emails. As a result of this, visitors of the legitimate pages and the resource owner will not recognize the presence of malicious content. Meanwhile, victims that were lured through malicious links will be affected directly.

Moreover, when harm is acknowledged or becomes visible to the intermediaries and

resource owners, taking action against it might not be as straightforward as one thinks. First of all, taking action against abuse or patching vulnerabilities has negative incentives such as the cost of infrastructure and abuse desks. This could easily raise the intermediaries' cost for security spending. In addition, resource owners might be required to pay for abuse and vulnerability remediation.

Another factor that plays a major role in not acting against the security problem is having a lack of technical knowledge to solve the problem. Generally, resource owners do not have the technical knowledge to act upon security notifications. Even when they know that their resource is insecure or used in malicious activities, they cannot perform the remediation steps themselves. As a result of this, they might have to bring in technicians who can solve issues to keep their resources secure. This cost and hassle might discourage resource owners to act against vulnerabilities and even malicious misuse of their resources.

Another negative impact is that unaware subscribers can see the protective countermeasures against abuse and vulnerabilities as a limitation to their Internet freedom. Some ISPs use walled garden notifications that place the infected customers' Internet connection into a quarantined environment where all Internet services are restricted. Thus, the resource owner's Internet experience will be interrupted to display the security notification. This type of proactive security measure might not be appreciated by the end users as their business or Internet experience will be disturbed until they remediate the problem. As a result of this disturbance, they might move to other intermediaries where no or less disruptive security measures are in use.

Furthermore, interventions against cybercrime might also affect legitimate resources and actions to collect intelligence from criminal infrastructure. For instance, to mitigate the Zeus malware threat, Microsoft performed several take-down actions, such as Operations b54/b71, to shut down the botnet command and control infrastructure. Microsoft relied on methods that are debated by the security community as they ended up hampering and even compromising several international investigations. Additionally, Microsoft operated on information that is devoted to tracking long-term cybercriminal activity. As a result of this, those operations diminished security industry tracking capabilities. In addition to this, dozens of legitimate domains were seized.

Not acting on the abuse notifications might lead to the degradation of services offered by the resource owner or the intermediaries. ISPs and hosting providers that do not act on the spammers in their network can be blacklisted by the bigger intermediaries and blacklist maintainers. As a result of this, the entire IP range that was used for legitimate reasons can be blacklisted and emails coming from these networks will be destroyed before reaching their destination. This will cause major disruption to email traffic in the network. If a resource owner does not act upon the security notifications, their resources can be blacklisted by third parties or the right to use their resources can be revoked by the intermediary. Blacklisting has critical consequences for businesses. For instance, when a compromised website is blacklisted by Google, they disappear from search results. Thus, the number of visitors to the website and therefore revenue of the business drops dramatically.

1.5 State of the art

In recent years, the effectiveness of abuse and vulnerability reporting that drives voluntary action has become a growing subject of research. In this section, we described prior research in 2 segments: (i) effectiveness of abuse notifications; and (ii) effectiveness of vulnerability notifications. Generally, prior research on abuse and vulnerability notifications investigated the effectiveness of the notifications in terms of vulnerability and abuse remediation.

1.5.1 Abuse notifications

Some researchers have assessed the effectiveness of abuse reporting and cleanup, often with the goal of understanding and improving the voluntary cleanup efforts. Various studies have explored how abuse notifications influence the cleanup of compromised servers and websites by using both direct and intermediary remediation strategies. When a legitimate server or website is compromised, often notifications are sent to the hosting provider that hosts the resource or owner of the website and asking them to clean the website. On the other hand, if the website is registered by malicious actors to be used in their malicious activities, the registrar and hosting provider is contacted and asked to take the website offline. In a prior study on abuse notification, Vasek et al. investigated the impact of verbose abuse notifications sent out to remediate compromised websites submitted to the StopBadware community feeds [32]. They randomly assigned compromised websites to three experimental groups: minimal notifications, detailed notifications that included all information from the minimal report and a more detailed description of the malware, and a control group where no notifications are made. For the minimal and detailed notification groups, they sent notifications to two entities: hosting provider and either website owners, for the compromised websites, or registrar if the website is registered by the malicious actors. Therefore, the study leveraged both direct and intermediary remediation strategies. They found that 62% of compromised websites assigned to a detailed notification group were cleaned within 16 days, compared to 45% of those assigned to minimal notifications group. Remarkably, they observe no difference in response rates between websites that are assigned to the control group and minimal notification group. This work showed the importance of providing detailed information about compromised when reporting to intermediaries and resource owners.

In an observational study, Li et al. investigated the impact of security notifications on over 700,000 compromised websites that were detected by Google Safe Browsing and Search Quality[33]. This study leveraged direct remediation strategies to promote cleanup. They found that security notifications sent via the Google Search Console promoted a 50% increase in the probability of cleanup. Furthermore, notifications reduced the duration of compromisation by 62%. In another study Canali et al. investigated hosting providers' ability to handle abuse [34]. As part of their research, they hosted their vulnerable websites on 22 hosting providers and repeatedly ran five different attacks on them that simulated

bot-like infection and then reported the hosting providers about these attacks on their test websites. Unlike other studies, the authors measured the effectiveness of intermediary remediation strategy. They found out that around 40% of the hosting providers deployed security mechanisms to block simple attacks and 36% of the hosting providers reacted to the abuse notifications. Hosting providers that responded to the reports only suspended the compromised websites. Additionally, the authors issued false abuse reports to measure the response to false positives. Surprisingly, they found out that 13% of the notified hosting providers took action based on the false abuse reports, despite a lack of evidence. This shows the possible pitfalls in the follow-up investigation on abuse reports. Most similarly, Nappa et al. issued abuse notifications to hosting providers that hosted 19 long-lived malware distributing websites [35]. Thus, the study leveraged an intermediary remediation approach to promote cleanup. Similarly to the previous study, only 39% of the hosting providers responded to the reports, taking an average of 4.3 days to take action.

Alternatively, abuse reports can be placed on websites that can be inspected by anyone. This approach sometimes has a positive impact on cleanup. This might be because abuse notifier might leverage this abuse data to promote direct and intermediary remediation. In a study on the lifetime of Zeus botnet C&C domains, Gañán et al. discovered that malicious domains displayed in public trackers were remediated more quickly than domains that were not reported and used for malware related intelligence gathering [36]. In another study, Tang et al. conducted a quasi-experiment publishing outgoing spam levels to change the behavior of the worst-performing network operators in countries with similar characteristics [37]. First, they mapped the spam data based on countries. Then, they assigned the countries with similar characteristics to two experimental groups: treatment group and control group. Spam data on countries in the treatment group is published on a website called spamranking.net. For countries in the control group, no notifications were made publicly or otherwise. Authors found that countries in the treatment group subjected to information disclosure reduced outgoing spam by approximately 16%.

Additionally, several studies investigated the effectiveness of sharing abuse data. Vasek et al. studied the effectiveness of sharing abuse data with proactive hosting providers [38]. In this case, providers approach an abuse and vulnerability database maintainer and ask for malicious links detected in their network to initiate the intermediary remediation process. The study observed the impact of sharing more than 28,000 malicious links which are shared with 41 hosting providers. Their results demonstrated that sharing has an immediate effect on cleaning the reported malicious links. However, they found out that long-lived abuse takes even longer to clean after being reported. In another study, Moore et al. found that refusing to share abuse data significantly slows down the cleanup efforts [39]. Moreover, Hutchins et al. provided evidence that expertise learned through abuse data sharing could increase the effectiveness of malicious website cleanup efforts [40].

1.5.2 Vulnerability notifications

A range of research has looked into the feasibility and efficacy of large-scale vulnerability reporting mechanisms. In one of the first studies, Durumeric et al. investigated how notifications to intermediaries can expedite vulnerability remediation [41]. To this end, they discovered servers vulnerable to a highly publicized OpenSSL vulnerability called Heartbleed. They notified intermediaries through the abuse email contact extracted from each WHOIS record to promote intermediary remediation. Their study discovered that when they notified network operators about the vulnerability in their network, the rate of patching increased by 47%. In another study, Kührer et al. worked on vulnerability notification campaigns for administrators of vulnerable Network Time Protocol (NTP) servers, in collaboration with CERTs and afflicted vendors [42]. Similarly to the previous study, authors prefer using an intermediary remediation approach. This was mainly because there was no scalable public contact information for the server owners. The authors reported a 92% reduction in vulnerable servers in under 3 months. While the study results are as impressive as it is, the study lacks a control group to assess the impact of notification campaigns for CERTs and device manufacturers.

Li et al. briefly investigated the impact of different aspects of vulnerability notifications that could play a role in terms of increasing vulnerability patching rates[43]. They studied who to send the notifications to and how much information needed to be included in the notification content. They mainly focused on intermediary remediation strategy by sending vulnerability notifications to hosting providers, ISPs and other organizations known to contact intermediaries to disseminate vulnerability and abuse data. Their findings demonstrated that vulnerability notifications addressed directly to the owners of the vulnerable network owners promote faster remediation than those sent to national CERTs and US-CERT. Besides this, their results also revealed that vulnerability remediation rates increased when network owners were contacted with detailed vulnerability notifications, compared to terse vulnerability notifications. On the other hand, their results showed that the majority of recipients did not take action or only partial remediation action was taken. Similarly, a study by Stock et al. measured the feasibility and effectiveness of large-scale notification campaigns for website and server vulnerabilities [44]. Their findings showed that only around 6% of the affected parties could be reached through notifications. Similarly, this study also reported low overall remediation rates. In a recent study, Stock et al. studied the effectiveness of other direct channels such as postal mail, social media, and phone to reach network and website owners [45]. Their study mainly relied on a direct remediation approach. They concluded that the slightly higher vulnerability remediation rates of these notifications channels do not justify the additional work and costs. More recently, Zeng et al. studied whether sending direct notifications to the owners of vulnerable sites could incentivize them to improve their misconfigurations [46]. Similar to previous studies, their results demonstrated a marginal but statistically significant effect on remediation. Lastly, Zhang et al. focused on remediating vulnerabilities in educational institution networks in China [47]. The study focused on promoting both direct and intermediary remediation

strategies. In their study, they measured the effectiveness of instant messaging (IM), telephone and email notifications. They determined that IM is the most effective notification method for such network settings.

1.6 Research Gaps

In recent years, various academic and industrial studies have been published to understand and address common problems in abuse and vulnerability reporting and remediation. Some researchers investigated the impact of the notification content on the effectiveness of voluntary action, while others focused on the feasibility of large-scale notifications. The security community mainly focused on providing recommendations and best practices to abuse desk employees so that they could address common security issues.

Prior work provided a foundation for understanding certain aspects of the voluntary action. On the other hand, we still know little about the factors and aspects that drive higher response rates to security notifications. This is mainly because various notification mechanisms and aspects of the notifications have never been researched systematically, only in limited specific instances such as spam blacklists or notification of phishing sites[48, 49].

Based on prior work, we identify three key gaps that this dissertation aims to investigate: understanding the impact of notification sender reputation on web-based malware cleanup, assessing who to notify and how to further incentivize to remediate web vulnerabilities and lastly identifying and improving effectiveness and issues of abuse and vulnerability notifications made by ISPs to their subscribers with infected or vulnerable machines.

First, we lack key empirical insights into the effectiveness of sender reputation on cleanup rates. Prior research investigates the effectiveness of abuse notifications without assessing the influence of sender reputation. As a result of this, we don't know whether their results are tied to the influence of the email addresses they used in their studies and if it is possible to increase the effectiveness of the abuse notifications by simply sending them from more reputable organizations.

Secondly, we lack evidence-based guidance on how to deliver the security-related information to the right hands and how to incentivize actors in acting against it. To our knowledge, there has been no work that studies the interaction between such notification mechanisms and the incentives of the affected intermediaries. Such research would help the security community identify actors with the strongest incentives to act upon the notifications and the most effective notification mechanisms to incentivize resource owners.

Finally, prior work did not study the effectiveness of existing voluntary efforts in broadband ISP networks. Typically, these remediation efforts leverage intermediary remediation strategies, as resource owners remediate the security issue after receiving a notification from their ISP. There are millions of infected and vulnerable resources in broadband ISP networks. It is crucially important to find out the effectiveness of currently available methods and ways to improve the effectiveness of these methods. For many of these resources,

there are no patches to remediate the vulnerabilities. Additionally, in many cases it is impossible to provide device-related information to resource owners. We don't know whether resource owners can act upon security notifications without device-specific cleanup advice. Moreover, we lack insight into potential issues experienced by notified parties in ISP networks. The perspective of the notification receivers is an understudied topic that could result in higher remediation rates if understood well enough. Currently, we have a very limited idea about why resource owners and intermediaries might choose not to act upon the notifications and what can we do to improve this. We need more empirical studies to identify and quantify the occurrence of these issues.

1.7 Research Aims and Questions

The main objective of this dissertation is to measure and increase the effectiveness of voluntary action against cybercrime. This objective requires us to systematically analyze the relationship between types of notification mechanisms and security issues at key Internet intermediaries, such as Internet service providers and hosting providers. Furthermore, this objective requires experiments with industry partners to measure findings on how to make notification mechanisms more effective. The main research question of this dissertation can be framed as follows:

How can the effectiveness of voluntary action against cybercrime be increased?

The main research question is further decomposed into five different studies. These studies and their findings are explained in the upcoming chapters. A brief introduction to these studies can be found below.

Study 1: Measuring the Role of Sender Reputation in Abuse Reporting and Cleanup

The first study deals with the impact of the reputation of the abuse notification sender. Not all reports are treated equally, as can be seen from the fact that some recipients assign a trusted status to some senders ('trusted complainer'), sometimes tied to a specific API for receiving the report and even semi-automatically acting upon it. However, does that make a measurable difference in terms of abuse remediation and cleanup?

The study aims to measure the role of the abuse notification sender's reputation by issuing technically similar abuse reports for compromised websites from various sources with different reputations. In this study, we used a private data feed of Asprox-infected websites to issue notifications from three senders with different reputations: an individual, a university and an established anti-malware organization. We compared their cleanup rates and speed to each other and a control group compromised with the same malware.

The study aims to answer the following questions:

- To what extent does sender reputation matter when notifying resource

owners and their intermediaries with evidence that their system is compromised?

- To what extent are abuse notifications effective in getting intermediaries and end users to act against abuse?

Study 2: Measuring the Impact of Large-scale Vulnerability Notifications Campaigns

In the second study, we focus on various aspects of large-scale vulnerability campaigns. The study mainly examines the impact of providing a mechanism to actively demonstrate the vulnerability in the content of the notification compared to static notifications without this mechanism. Additionally, the study investigates the incentives of the different affected parties. Based on that, the study wanted to identify the most effective actor to contact.

In this study we investigated the following two research questions:

- What communication path mobilizes the strongest incentive for remediation; delivering security notifications directly to nameserver operators, their customer or the network operator?
- What is the impact of providing recipients a mechanism to actively demonstrate the vulnerability for their own system, rather than sending them the standard static notification message?

Study 3: Evaluating the Effectiveness of Quarantining Compromised Users in Walled Gardens

The third study investigated the usability and effectiveness of quarantining and walled garden notifications in terms of aiding users in residential networks to clean up malware-infected machines. The study observed 1,736 quarantining actions involving 1,208 subscribers of a medium-sized ISP. Each one of these subscribers had a malware infected device in their houses that need to be remediated. The study explored the impact of three mechanisms to release users from the quarantine environment, infection type and the time end users spend in quarantine on cleanup in great detail. Additionally, the study briefly presented the actual experience of the quarantined end users by analyzing the communication between ISP and the quarantined end users.

In short, the study explores the following research questions:

- To what extent are walled garden notifications effective at getting end users to remediate malware infection in residential networks?
- How much pushback do ISPs face from their quarantined users?

Study 4: Evaluating the Effectiveness of ISP-made vulnerability notifications

After evaluating the impact of walled garden notifications on malware-infected devices, we assessed the effectiveness of walled garden notifications on vulnerable devices in residential networks. The study measured the remediation rates achieved by a medium-sized ISP for 1,688 retail customers running open DNS resolvers or Multicast DNS services. These devices had the potential to be used in UDP-based amplification attacks. The study evaluated the effectiveness of the walled garden notifications by comparing them it to email notifications and natural remediation. The study also provided explanations for surprisingly high natural remediation rates. Moreover, the experiences of the notified users are presented in great detail.

The study provides answers to the following research questions:

- To what extent are walled garden notifications effective at remediating vulnerable devices in residential networks compared to email notifications?
- What are the issues raised by recipients of ISP-made vulnerability notifications?

Study 5: Evaluating the effectiveness of ISP-made notifications to users with compromised IoT devices

In the last study of this dissertation, we investigated IoT malware cleanup in the network of a medium-sized ISP. To measure IoT malware remediation rates, we combined data from an observational study and a randomized controlled trial involving 220 subscribers who were infected with Mirai IoT malware together with data from honeypots and darknets. The observational study measures the effectiveness of the existing ISP walled garden mechanism used by the ISP. Randomized controlled experiments assessed the impact of the walled garden and email notifications with improved content tailored to IoT infection remediation compared to a control group where infected device owners were not notified. Additionally, customer experiences and actions were analyzed via 76 phone interviews and the communications logs of the ISP.

In short, the following research questions are answered:

- What is the most effective method to notify the end users against compromised IoT devices?
- How did the end users react to the IoT malware notifications made by their ISP?

1.8 Dissertation Outline

The structure of this dissertation is as follows. Chapters 2 through 6 explain the five peer-reviewed studies introduced in 1.7. Lastly, in Chapter 7 the recaps, conclusions and future work of this research project are described.

Empirical research presented in each chapter has been published in journals and peer-reviewed venues. Table 1.2 shows cybersecurity researchers that I was fortunate enough to collaborate with while conducting these studies. It is my pleasure to explain their valuable contributions to each study in Appendix F.

Table 1.2: Outline of dissertation chapters 2 to 6

Chapter Number	Publication
2	O. Cetin, M. Jhaveri, C. Gañán, M. van Eeten, and T. Moore, "Understanding the role of sender reputation in abuse reporting and cleanup" In Workshop on the Economy of Information Security (WEIS),2015. O. Cetin, M. Jhaveri, C. Gañán, M. van Eeten, and T. Moore, "Understanding the role of sender reputation in abuse reporting and cleanup" Journal of Cybersecurity 2, no. 1 (2016): 83-98.
3	O. Cetin, C. Ganán, M. Korczynski, and M. van Eeten, "Make notifications great again: learning how to notify in the age of large-scale vulnerability scanning", In Workshop on the Economy of Information Security (WEIS),2017.
4	O. Çetin, C. Gañán, L. Altena and M. van Eeten, "Let me out! evaluating the effectiveness of quarantining compromised users in walled gardens." In Fourteenth Symposium on Usable Privacy and Security (SOUPS) 2018, pp. 251-263. 2018.
5	O. Çetin, C. Gañán, L. Altena, S. Tajalizadehkhoob, and M. van Eeten, "Tell Me You Fixed It: Evaluating Vulnerability Notifications via Quarantine Networks" In 2019 IEEE European Symposium on Security and Privacy (EuroS&P).
6	O. Çetin, C. Gañán, L. Altena, T. Kasama, D. Inoue, K. Tamiya, Y. Tie, K. Yoshioka, and M. van Eeten, "Cleaning Up the Internet of Evil Things: Real-World Evidence on ISP and Consumer Efforts to Remove Mirai", in 2019 Network and Distributed System Security Symposium (NDSS) 2019.

CHAPTER 2

Measuring the effectiveness of abuse notifications made to hosting providers

Participants on the front lines of abuse reporting have a variety of options to notify intermediaries and resource owners about the abuse of their systems and services. These can include emails to personal messages to blacklists to machine-generated feeds. Recipients of these reports have to voluntarily act on this information. We know remarkably little about the factors that drive higher response rates to abuse reports. One such factor is the reputation of the sender. In this chapter, we present a study that measures the impact of abuse notifications and notification sender's reputation on compromised cleanup rates. In the first part of the study, we measured the effectiveness of the abuse notifications by comparing a group of compromised websites without any notifications. In the second part of the study, we assess the effectiveness of issuing notifications from three senders with different reputations: an individual, a university and an established anti-malware organization. Additionally, we also studied the efficacy of cleanup advice provided via a link in the notifications.

2.1 Introduction

Advances in detecting and predicting malicious activity on the Internet, impressive as they are, tend to obscure a humbling question: Who is actually acting against these abusive resources? The reality is that the bulk of the fight against criminal activity depends critically on the voluntary actions of many thousands of providers and resource owners who receive abuse reports. These reports relay that a resource under their control – be it a machine, account, or service – has been observed in malicious activity. Each day, millions of abuse reports are sent out across the Internet via a variety of mechanisms, from personal messages to emails to public trackers to queryable blacklists with thousands of hacked sites or millions of spambots.

Proactive participants may pull data from clearinghouses such as Spamhaus and Shadowserver. But in many cases, the reports are pushed to recipients based upon publicly available abuse contact information. In these circumstances, those who can act against the

abusive resource might never actually see the information. If the information does reach them, it might be ignored, misunderstood or assigned low priority. Still, against all these odds, many reports are acted upon, without any formal requirement, across different jurisdictions and often without a pre-established relationship between sender and recipient. This voluntary action is an under-appreciated component of the fight against cybercrime.

Remarkably little research has been undertaken into what factors drive the chances of a recipient acting upon an abuse report (notable exceptions are [32, 34, 41, 42]). One factor, the reputation of the sender, clearly plays an important role in practice. Not all reports are treated equal, as can be seen from the fact that some recipients assign a trusted status to some senders ('trusted complainer'), sometimes tied to a specific API for receiving the report and even semi-automatically acting upon it.

The underlying issue is a signaling problem, and therefore, an economic one. There is no central authority that clears which notifications are valid and merit the attention of the intermediary or resource owner. This problem is exacerbated by the fact that many intermediaries receive thousands of reports each day. One way to triage this influx of requests for action is to judge the reputation of the sender.

We present the first randomized controlled experiment to measure the effect of sender reputation on cleanup rates and speed. During two campaigns over December 2014–February 2015, we sent out a total of 480 abuse reports to hosting providers and website owners from three senders with varying reputation signals. We compared their cleanup rates to each other and to a control group compromised with the same malware.

In the next section, we outline the experimental design. In Section 2.3, we turn to the process of data collection, most notably tracking the cleanup of the compromised resources that were being reported on. The results of the experiment are discussed in Section 2.4. Surprisingly, we find no evidence that sender reputation improves cleanup. We find that the evasiveness of the attacker in hiding compromise can substantially hamper cleanup efforts. Furthermore, we find that the minority of hosting providers who viewed our cleanup advice were much more likely to remediate infections than those who did not, but that website owners who viewed the advice fared no better. We compare our findings to related work in the area in Section 2.5. We describe limitations in Section 2.6 and conclude in Section 2.7.

2.2 Experimental Design

Does sender reputation matter when notifying domain owners and their hosting providers with evidence that their website is compromised? We designed an experiment measuring cleanup rates as a result of abuse reports sent from three senders with varying levels of reputation: an unknown individual, a university and StopBadware, a well-established non-profit organization that fights malware in collaboration with industry partners.

The analysis and data collection started in December 2014 and continued through the first week of February 2015 across two campaigns. Figure 2.1 illustrates the rules we

applied to get the experimental data set from the original feed.

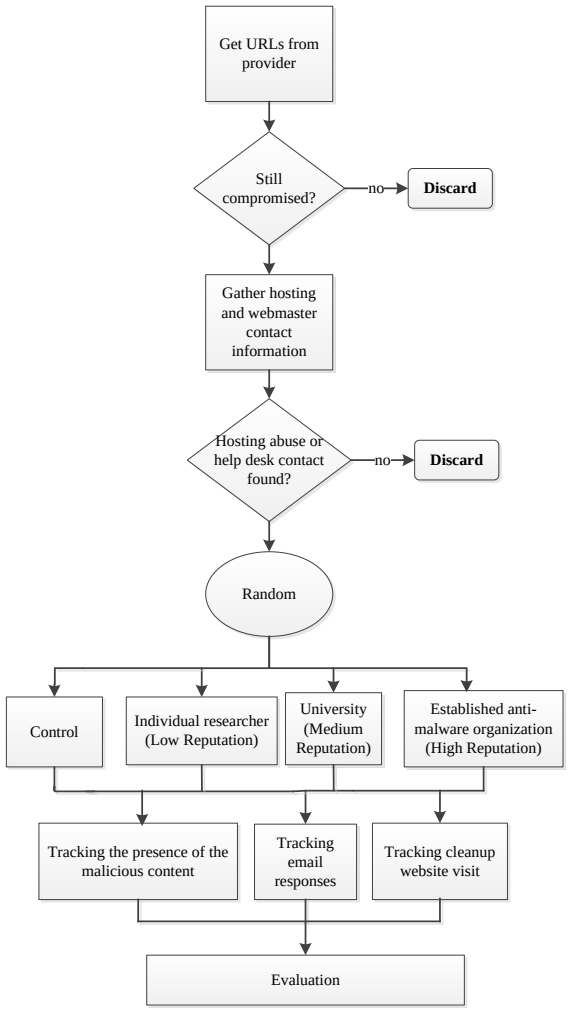


Figure 2.1: Flow diagram of the progress through the phases of our experiment

2.2.1 Study Population and Sampling

The study population was derived from a raw daily feed of URLs serving malicious downloads originating from the Asprox botnet. This private source of abuse data was not shared with anyone else and free of any prior notification attempts.

From December 7th, 2014 until January 19th, 2015, we received a total of 7,013 infected URLs. The daily feed fluctuated dramatically, with peaks of close to one thousand URLs and days with just a handful. Most days, we received between 50-100 URLs. From these, we took a daily random sample, typically of around 40 URLs. We could not include all URLs we received in the experiment because of a bottleneck further on in the process: tracking the up-time of the compromised content (see Section 2.3).

We issued notifications within a day of first reporting. Before reporting, we checked whether the reported site was indeed still compromised. In a handful of cases, cleanup or remediation seemed to have taken place already. If so, the URL was discarded. Next, we looked up abuse contact information for the hosting provider and the the domain owner from WHOIS data. If we could not find any contact information for the hosting provider (for example, if the WHOIS information was set to private), we discarded the URL. When we did not find any contact information for the domain owner, we would use the RFC standard abuse e-mail address [50]. All in all, we discarded fewer than 10 URLs for either no longer being compromised or the lack of an abuse contact for the hosting provider.

To determine the total sample size, in other words how many URLs we needed, we completed a power calculation for the main outcome variable, cleanup rate. We estimated power for three levels: 80%, 85% and 90% and used a 5.65 standard deviation based on prior studies [32]. Differences in mean sixteen-day cleanup time of about 0.84 days between conditions can be detected with 90% power in two-tailed tests with 95% confidence, based on a sample of 80 websites in each treatment group. To ensure that the control has enough statistical power for baseline comparison across treatment groups, we set the control equal to all other treatment groups combined. This resulted in a total sample size of 480 URLs. However, URLs were distributed over two campaigns. For each individual campaign, we did not meet power analysis conditions. That being said, we found significant differences for the second campaign (see Section 2.4).

2.2.2 Treatment Groups & Rationale

Using a random number generator, we assigned URLs to a treatment condition or to the control group. The three treatment conditions were sending an abuse report from an individual researcher, a university and an established anti-malware organization (see Table 2.1). The report from the individual researcher was designed to reflect a low reputation abuse notifier and was sent from a Gmail account. The university group was set up to reflect a medium reputation abuse notifier. Here, we used a functional e-mail address from Delft University of Technology. The established anti-malware organization was included as the sender with

the highest reputation. StopBadware generously provided us an e-mail account at their domain to send notifications on their behalf [51].

Table 2.1: Overview of each treatment group

Group	Description	E-mail Address	Sample Size Camp. 1	Sample Size Camp. 2	Rationale
Control	No Notification	N/A	17	229	Baseline to understand the natural rate of compromised host survival
Individual researcher	Individual internet researcher	malwarereporting@gmail.com	23	57	Individuals may send mixed signals, from quality to motivation
University	Academic Institution	malwarereporter-tbm@tudelft.nl	17	62	Academic organizations may signal higher quality and research intent
Established Anti-malware Organization	Anti-malware nonprofit organization	abuse-reporter@stopbadware.org	20	61	Dedicated organizations may signal the highest quality research and/or potential commercial enforcement

As the randomization took place at a URL level, the domain owner and the hosting provider were assigned to the same treatment group. The notified entities were, by nature of the intervention, not blinded.

Once assigned, we completed a statistical analysis on key attributes to ensure the assignments were comparable across groups. The control group served as a baseline to understand the natural survival rate of a compromise and was the only one not to receive notifications. There was no difference among the treatment groups other than the domain of the e-mail address and the host of the cleanup content. We base this on studies (e.g. [52]) that indicate users perceive domains with certain top-level extensions to have differing levels of authority in terms of the accuracy of information.

2.2.3 Notification & Cleanup Support Site

The abuse notifications were based on the best practice for reporting malware URLs that has been developed by StopBadware [53]. The content included the malicious URL, a descrip-

tion of the Asprox malware, the IP address, date and time of the malware detection and a detailed description of the malware behavior. Abuse notification sample for established anti-malware organization, university and individual internet researcher are respectively presented in appendix (see figure A.1, A.2 and A.3).

We sent notifications to each treatment group during 12 days in total. All treatment groups received an identical abuse notification, except for the sender e-mail address and the included link to a web page where we described cleanup advice for sites compromised by Asprox. The web page provided a brief guide explaining how to identify and remove Asprox malware and backdoors from compromised websites. The page also included links to other websites for precautionary measures to prevent the site from being compromised again. Figure A.4, A.5 and A.6 in the appendix, contains samples of the various cleanup websites shared in the e-mail notification for each of the treatment groups.

The webpage was hosted at different domains consistent with each treatment condition. The individual researcher e-mailed a link to a free hosting webpage, the university to a page inside the official TU Delft website, and StopBadware to a page on their official domain.

Furthermore, each cleanup link contained a unique seven-character code allowing us to track which recipients clicked on the link. In this way, we measure whether visiting the cleanup page was associated with higher cleanup rates.

To prevent biases because of the recipients' varying abilities to receive the e-mail and view the webpage, we tested all the e-mail notifications across various e-mail services to ensure correct delivery and double-checked that the webpages were not on any of the major blacklists.

2.2.4 Evaluation

We evaluate the experiment based on the differences in cleanup rates and median-time to cleanup across the various treatment groups relative to the control group. We also explore the relationship between cleanup rates and other variables, such as visits to the cleanup advice page and the responses of providers to our notifications.

2.3 Data Collection

To perform the experiment designed in the previous section, we received assistance from an individual participating in the working group analyzing and fighting the Asprox botnet. He supplied us with a private feed of URLs in use by Asprox. The URLs were captured via spamtraps and various honeypot servers located in Europe and the United States.

The Asprox botnet was first detected in 2007. Since then, it has evolved several times. Currently it is mostly used for spam, phishing, the distribution of malware to increase the size of its network, and for the delivery payload of pay-per-install affiliates [54]. Asprox

compromises websites by building a target list of vulnerable domains and then injects SQL code that inserts a PHP script that will trigger the visitor to download malware or redirect them to various phishing sites. Our URL feed contained both variations.

2.3.1 Evolution of Asprox compromised sites

In the course of our experiment, Asprox's behavior changed as it went through two different attack campaigns (see Table 2.2). From December 2014 until beginning of January 2015, the infected sites delivered a malicious file. After that, from January 2015 until February 2015, instead of delivering a malicious file, infected domains redirected visitors to an ad-fraud related site. Moreover, these two campaigns did not only differ on the type of malicious behavior but also on the countermeasures taken by the botnet against detection and removal.

Table 2.2: Overview of each campaign

Campaigns	Start Date	End Date	Type	Character
Campaign 1	12/08/2014	12/26/2014	Malware	* Customized and standard error messages * IP and identifier based blacklisting
Campaign 2	01/12/2015	02/04/2015	Ad-fraud	* Standard error message

During the first campaign, the botnet's countermeasures included blacklisting of visitors to the compromised sites based on IP addresses and machine fingerprinting. The blacklist was managed by back-end command-and-control systems and shared among the compromised sites.

Once an IP address was blacklisted, the compromised sites stopped serving the malicious ZIP file to that particular IP and displayed an error message instead. We encountered two different types of error messages: (i) HTTP standard error messages such as 404 Not Found, and (ii) customized error messages such as "You have exceeded the maximum number of downloads". In addition, sites only accepted requests coming from Internet Explorer 7 and versions above.

In contrast to the first campaign, the second campaign did not apply any type of blacklisting. Instead the main countermeasure consisted of displaying an error message when trying to access the malicious PHP file alone. Moreover, the path to reach the malicious content would change periodically.

In most cases, the malicious content was only accessible through the URLs included in the phishing e-mails. These URLs included a request code that allowed infected sites to serve malware binaries and phishing pages that belonged to a specific Asprox attack. Once that specific attack ended, the compromised sites stopped responding to the corresponding URLs and displayed an error message instead. Table 2.3 shows a list of request codes and

the corresponding attributes for both malware and phishing URLs. For instance, "?pizza=" code was only used for triggering PizzaHut_Coupon.exe Asprox malware binary.

Table 2.3: Examples request codes and what they represent.

Malware Campaign			
Request Code	*Targeted Companies*	*Sample*	*Name of Executable*
?c=	Costco	?c=r24t/fwI8nYJeoktSMii3IkC8ItN3Dqcpphcm375Sg4	Costco_OrderID.exe
?fb=	Facebook	?fb=i2uXy5/kOZ77bjvMAA0hgsai4YbZNvC78Ji7amd1D8Y	FB-Password-Reset_Form.exe
?w=	Walgreens	?w=uhUGpftxxueBCfO/6FxAx7p2/Guz9BjRwRj/1YVMcKI	Walgreens_OrderID.exe
?pizza=	Pizza Hut	?pizza=Wa5wEaLOSojFl3kTaW3OIgOW150DCm7Jda8m83pzVJo	PizzaHut_Coupon.exe
Ad Fraud and Phishing Campaign			
Request Code	*Type of Scam*	*Sample*	
?po=	Ad-Fraud	?po=rIdsS+cFDm7bNp4duz57G0IWqGTH15cqcKUdvtSGBME	
?r=	Dating Website Scam	?r=2	

2.3.2 Tracking presence of malicious content

Given the evolution and countermeasures of the Asprox botnet, the experiment required a complex methodology to track the notified entities acted upon our abuse report and cleaned up the compromised site. In the following, we describe the notification process and the methodology to track Asprox infected websites.

To identify and monitor malicious content for the first campaign, we first required a mechanism to bypass the botnet's blacklisting of visitors based on IP addresses and fingerprinting. The compromised sites used error messages to make it harder to distinguish malicious links from broken or dead links. We developed an automated tool that used IP addresses from 2 private and 7 public HTTP proxy services and checked whether the IP address that the tracking tool received had not been used before. Each day, 3 different proxy services were selected. All new IP addresses were checked against a list of previously used IP addresses. If it has been previously used, we discarded it. If not, we added it to the list. The IP addresses were selected following a round-robin algorithm from the pool of proxy services.

During a 16-day tracking period, we followed the procedure outlined in Figure 2.2 to determine whether a site was considered to be clean or compromised. Exactly 16 of the 486 total compromised sites (3%) periodically did not resolve. All were from the second campaign: 10 in the control group, 4 in the established anti-malware organization group, and 2 in the individual researcher group. While this might imply the site has been cleaned, that isn't always the case. Earlier work indicates that clean-up actions are sometimes visible in the WHOIS data [32], specifically in the status fields. We identified three cases (two in established anti-malware organization group and one in individual researcher group) where

the Domain Status and other fields of the WHOIS records changed, indicating that content of the site was removed. In the other 13 cases, we had no clues to clearly determine whether the site was actually cleaned up or in temporarily maintenance. Thus, we considered these 13 cases still infected.

Finally, in situations where the domain name resolved but the URL returned an HTTP error code different from HTTP 404 (Not Found), we also assumed that the malicious file was still present.

When a server successfully returned some content or a redirection to another website, our scanner analyzed the content searching for common Asprox malicious behavior. This procedure is summarized in Figure 2.3.

In both campaigns, we started by accessing the infected website and analyzing the HTTP server header request. If the server returned HTTP 200 (OK), then we further analyzed the header's content-disposition field to assess the attachment of a file with a .zip extension, which would contain the malicious binaries. If the website delivered a zip file, we concluded that the malicious script was still present and the website remained compromised.

The absence of an attachment in the website did not necessarily indicate that the site was clean. In some cases, infected sites were acting as redirectors to various phishing and ad-fraud sites. To capture this behavior, we analyzed the HTML content of the infected websites looking for a specific combination of HTML tags that were used for redirecting to known ad-fraud and rogue pharmacy sites that were captured during previous scans. If the redirected site led to malicious content we marked it as being compromised.

When clearly malicious content was not present in the redirected site, we manually entered it into the VirusTotal [55] website query field. We then selected "Re-Analyze" to force the service to check whether the site was blacklisted at that time. When the site returned that the URL or domain was in the blacklist, we marked it as being malicious. When indicated as being clear, we followed up and ran it through VirusTotal's passive DNS replication service to see if the resolved IP address hosted any other Asprox-related site. If found, we concluded that the site was still compromised.

When conditions were unclear whether malicious file is removed, we consider sites still malicious. These conditions include PHP fatal errors, disabled, and suspended pages. Disabled and suspended pages might indicate that action was taken to mitigate the abuse, even though the malicious script might still remain. In two cases, malicious links displayed a PHP fatal error [56]. While this could be related to a programming error, the ones we reviewed included HTML tags that are specifically associated with malicious content. Hence we assume that this implied the site was still compromised, and possibly just temporarily generating the fatal error to hide from hosting provider clean-up efforts.

When the website returned a HTTP 404 (Not Found) error message or in the absence of a clear indicator of malicious content, we classified the compromised site as potentially clean since the botnet infrastructure had modules to prevent security bots from reaching the malicious content. To gather more information about these potentially clean websites,

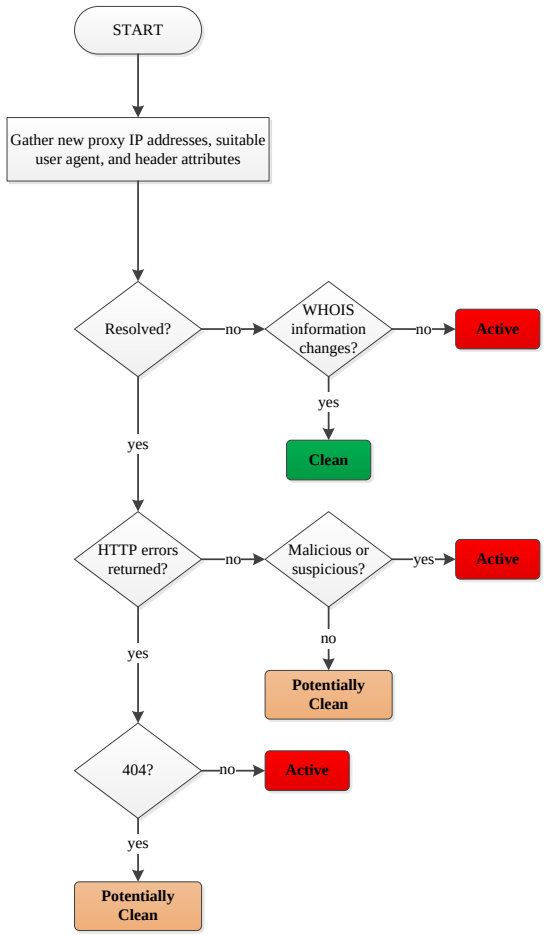


Figure 2.2: Flow chart for following up to determine when clean

we scan those sites 2 more times on the same day. If during these 2 additional scans no indicators of malicious or suspicious behavior were found, follow-ups scans were performed during the next 2 days with 3 unique requests. If there was no malicious or suspicious behavior during 3 consecutive days, then we considered the site to be potentially clean and manually investigated the URLs using online server header checker websites (e.g. [57]) and

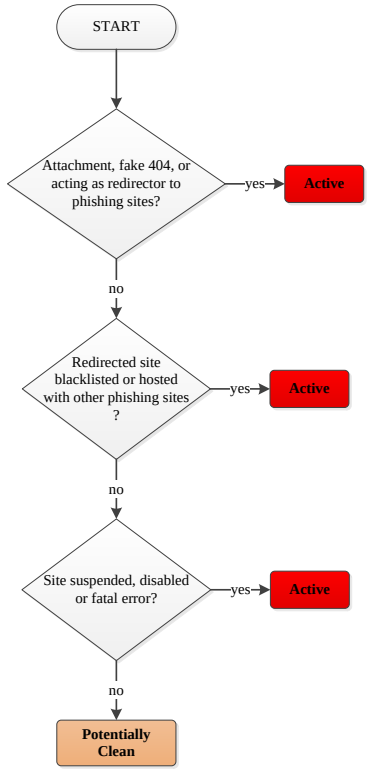


Figure 2.3: Flow chart for deciding whether a site is malicious

by visiting it manually using a 'clean' set of IP addresses that were acquired via a premium VPN subscription. These manual follow-ups were made to ensure reliable measurements on the presence of malicious content. The evolution of Asprox made it impossible to fully rely on automation. In the end, we only considered a site clean if it was never subsequently observed to be malicious in manual and automated scans.

During the second campaign, the botnet infrastructure was no longer using blacklisting based on IP addresses or fingerprinting. Therefore, we only used IP addresses from a single HTTP proxy service to track the presence of malicious content. As a preventive measure, our scanner used a mechanism where IP addresses were changed twice a day and different browser suits were used to visit the site. Only one followup was made for each day of tracking due to lack of blacklisting. Another difference with the first campaign was that

scans for the last day of tracking was automated. We only considered a site clean if, and only if, there was no malicious content related to Asprox botnet in both followups and last day scans.

Throughout the tracking process of the second campaign, compromised sites stopped redirecting to ad-fraud sites and paths to ad-fraud campaign were displaying standard error messages. This indicated that Asprox ad-fraud campaign was over. New links were generated by the botmasters for redirecting to the new scams sites such as fake dating or diet websites. Thus, the same infected websites that were used during the second campaign to redirect to ad-fraud related websites were now being used to redirect to other type of scams.

2.3.3 Tracking affected party responses

As part of the experiment, we also regularly checked the inbox of the different e-mail accounts created for this study. We received automated and manual responses from the affected parties. Automated responses came from hosting providers to acknowledge the reception of our notification. Most of the automated responses contained a ticket number, to be included in further communication about the infection. Some providers also included details of the ticket along with a URL for tracking the incident status.

For abuse notification we issued to CloudFlare, we received automated responses mentioning the abuse contact information for the hosting provider. However, we did not take any additional step because CloudFlare forwarded our notifications to site owners and the hosting provider.

Manual responses came from domain owners and abuse-desk employees to inform us about the cleanup action taken or requesting more evidence about the compromise. When we received a manual response stating that appropriate action was taken, we re-scan the website to confirm this action. If the results of the scan found that the infection was still present, we responded to the corresponding entity stating the existence of the malicious PHP script. In these responses, a HTTP header request from the malicious URL was included to serve as evidence showing the existence of the malicious file. When more evidence of the compromised was requested, a brief explanation of the compromise and a specific solution was given.

We also analyzed the logs of our web pages with cleanup advice. Via the unique codes included in the URLs, we identified which hosting provider or site owner visited one of our cleanup websites. Unfortunately, we discovered in the course of the experiment that the server logs for the StopBadware page could not be analyzed, as the webserver relied on Cloudflare's CDN service to serve the static content, thus leaving no log of the visit [58].

2.4 Results

From December 7th, 2014 until January 19th, 2015, a total of 7,013 infected URLs were identified. From these we excluded less than 10 URLs that were not active or for which we were not able to obtain reliable contact information for the hosting provider. The daily feed fluctuated dramatically, with peaks of close to one thousand URLs and days with just a handful. Most days, we received between 50-100 URLs. From these, we took a daily random sample, typically around 40. Over time, this accumulated to a random sample of 486 URLs.

In the following we empirically estimate the survival probabilities using the Kaplan-Meier method. Survival functions measure the fraction of URLs that remain infected after a period of time. Because some websites remain infected at the end of the study, we cannot directly measure this probability but must estimate it instead. Differences between treatment groups were evaluated using the log-rank test. Additionally, a Cox proportional regression model was used to obtain the hazard ratios (HR). All two-sided p values less than 0.05 were considered significant.

2.4.1 Measuring the impact of notices

First, we determined whether sending notices to hosting providers and domain owners had an impact on the cleanup of the infected URLs. Table 2.4 provides some summary statistics regarding the status of the infected URLs 16 days after the notification. Entries are given for each treatment group. We reported the percentage of websites that were clean and the median number of days required to clean up those sites.

Table 2.4: Summary statistics on the time to clean up, according to the treatment group

Treatment type	Campaign 1			Campaign 2		
	#	% Clean	Median clean up time	#	% Clean	Median clean up time
Control	17	35.29%	14 days	229	26.20%	8 days
Indiv. researcher	23	69.57%	4 days	57	49.12%	2.5 days
University	17	64.71%	4 days	61	44.26%	3 days
Anti-malware Org.	20	80.95%	2 days	62	48.39%	1.5 days

It is worth noting the significant difference between the two malware campaigns that took placed during our experiment. From table 2.4, we can see that while 35% of the websites in the control group were clean after 16 days during the first campaign, only 26% of the websites in the control groups during the second campaign remediated their

infection. The same trend was observed for the rest of the treatment groups, i.e., lower cleanup rates were achieved during the second campaign than during the first campaign. For instance, the percentage of remediated infections for the high-reputation group was reduced from 81% in the first campaign to 49% in the second campaign. We attribute these differences to the behavior change of the Asprox botnet which became harder to identify and remove during the second campaign (see Section 2.3).

To further investigate whether these differences are significant, we compute the survival probabilities for each of the two different campaigns. Figure 2.4 plots these curves. This figure shows that 36% of websites that were notified during the first campaign remained infected after 16 days, compared to 65% for those that were notified during the second campaign. The log-rank test corroborated that the cleanup rate was significantly different during the two campaigns ($\chi^2 = 21.39, p = 3.75e-06$). Proportional hazard model was used to compute the adjusted-hazard ratio (HR) for the two campaigns with 95% confidence intervals (CI). The HR for remediating the infection in the first campaign was 2.11 (95%CI, 1.52-2.89) versus the second campaign, i.e., infected domains in the first campaign were cleaned up 2 times faster than during the second campaign. As both campaigns had significantly different cleanup rates, in the following we analyze them separately.

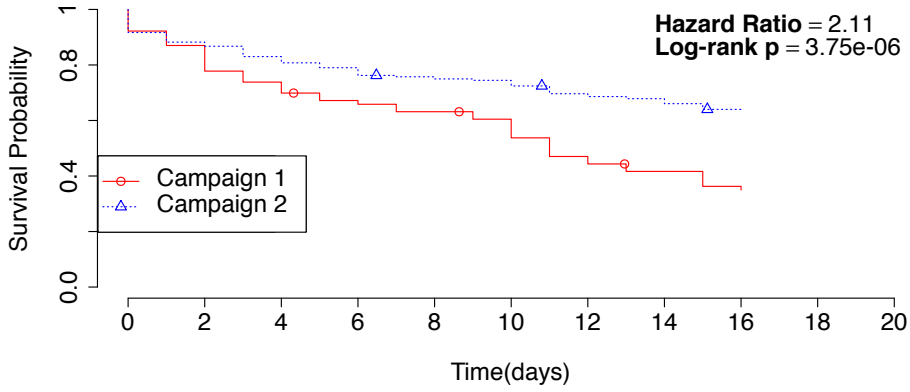


Figure 2.4: Survival probabilities for each notification campaign. The overall cleanup rates are lower in the second campaign when infections were harder to verify by providers.

Campaign 1

Comparing the percentage of clean websites of the control group with the other treatment groups, we can estimate whether the notices made a difference in terms of expediting the cleanup. As shown in Table 2.4, the control group always achieved a lower percentage of clean websites than the other groups. For instance, the median number of days to clean an Asprox-infected website was 14 days when no notice was sent. However, the median number of days to remediate an infection was greatly reduced when notices were sent. Websites in the high-reputation group were cleaned after 4 days in average. This supports the hypothesis that notices expedite the cleanup process.

Table 2.5: Log-rank test results (Campaign 1)

Group	Control		Indiv. researcher		University		Anti-malware Org.	
	$\tilde{\chi}^2$	p-value	$\tilde{\chi}^2$	p-value	$\tilde{\chi}^2$	p-value	$\tilde{\chi}^2$	p-value
Control			8.2	0.0041	6	0.0139	17.1	0.00003
Indiv. researcher	8.2	0.0041			0.2	0.644	1.7	0.198
University	6	0.0139	0.2	0.644			2.8	0.0972
Anti-malware Org.	17.1	0.00003	1.7	0.198	2.8	0.0972		

Again, to assess whether these difference are significant, we compute the survival probabilities for the different treatment groups (see Figure 2.5). We can observe different cleanup rates between the control group and the treatment groups which received notices. This figure shows that 65% of websites that were not notified remained infected after 16 days, compared to 30%, 35%, and 19% for those that belonged to the low-reputation, medium-reputation and high-reputation group respectively. The log-rank test confirms that these differences between the groups that received notices and the control group are significant ($\chi^2 = 15.61, p = 0.0014$). However, the differences among any of treatment groups which received notifications are not significant (see Table 2.5).

Campaign 2

In the previous section, we analyzed the impact of the notices that were sent during the first campaign and proved that sending notices expedited the cleanup process. In the following, we analyzed the impact of the notices sent during the second campaign that took place during January 2015.

As shown in Table 2.4, during this second campaign the percentage of sites successfully remediated was lower than during the first campaign. The control group had the lowest percentage of remediated infections, i.e., only 26% of websites were cleaned up. The rest of treatment groups achieved similar percentage of remediated sites (44%-49%). Therefore, though notices did impact the cleanup process, the reputation of the sender did not significantly affect that process.

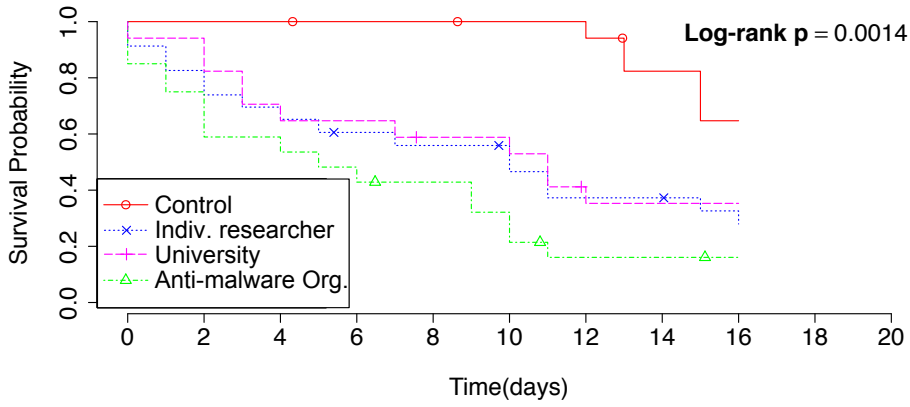


Figure 2.5: Survival probabilities per treatment group (Campaign 1)

Despite having a lower overall cleanup ratio, the sites that were remediated during the second campaign were cleaned up faster than in the first campaign. The median number of days before cleanup took place was 4 days during the second campaign, while it took 11 days during the first campaign. This suggests that the Asprox infections during the second campaign were harder to identify, but when detection was successful, clean up was done faster.

A plausible explanation for this pattern is to see it as the outcome of competency of the hosting provider. Those that are willing and able to recognize the compromise are also the ones that will be faster in terms of doing cleanup. Those that are not willing and able, will be slower in cleaning up or not do it at all. This explanation is consistent with the differences in cleanup between the two campaigns: at that time the malicious files of Asprox were easier to uncover, more hosting providers were able to initiate cleanup, including the less competent ones. The latter are likely to act more slowly, raising the median cleanup time.

We compute the survival curves for this second campaign per treatment group. Figure 2.6 plots the Kaplan-Meier estimates. In this campaign, the similarity among the treatment groups that received notices is even more clear than in the first campaign. This figure shows that after 5 days after tracking begun, 90% of websites that were not notified remained infected, compared to 64%, 63% and 65% for those that belonged to the low-reputation, medium-reputation and high-reputation group respectively. The log-rank test confirms that these differences between the treatment groups and the control group are significant ($\chi^2 = 28.39, p = 3.01e-06$). However, the differences among any of treatment groups are not significant (see Table 2.6).

Therefore, though the notices were effective during both campaigns, the clean-up rates

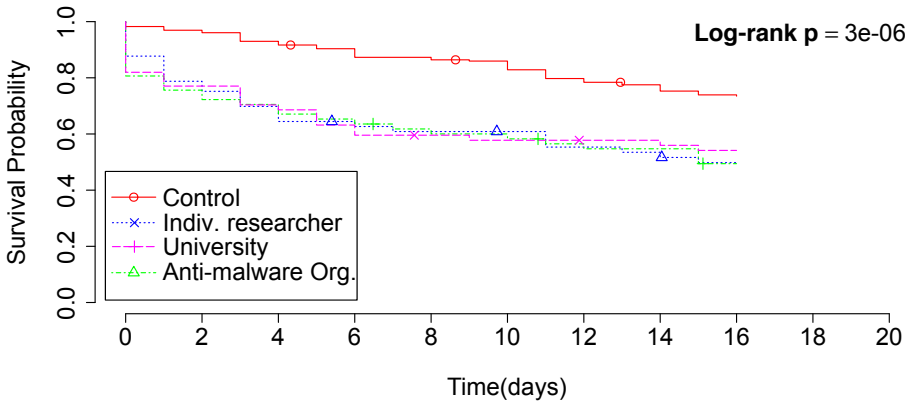


Figure 2.6: Survival probabilities per treatment group (Campaign 2)

Table 2.6: Log-rank test results (Campaign 2)

Group	Control		Indiv. researcher		University		Anti-malware Org.	
	$\tilde{\chi}^2$	p-value	$\tilde{\chi}^2$	p-value	$\tilde{\chi}^2$	p-value	$\tilde{\chi}^2$	p-value
Control			17.1	3.51e-05	13.6	22.1e-05	18.8	1.43e-05
Indiv. researcher	17.1	3.51e-05			0.1	0.746	0	0.919
University	13.6	22.1e-05	0.1	0.746			0.2	0.678
Anti-malware Org.	18.8	1.43e-05	0	0.91	0.2	0.678		

were higher during the first campaign. In neither of the campaigns did we observe a significant impact of sender reputation.

2.4.2 Efficacy of the clean-up advice websites

As part of the experiment, we created three websites to assist the cleanup process. The corresponding link to these website was included in the abuse report. As it turns out, few recipients clicked the link. During the 16-day follow-up, we tracked the visitors to the web pages at the university and the free hosting site.[1] The number of visitors is presented in Table 2.7. As can be seen, only 8.97% of the hosting providers visited our cleanup website. Similarly, only 7.48% of the contacted website owners visited our cleanup website.

To analyze if of the cleanup websites did help expedite remediation, we measure the

[1]We were unable to track the visitors of the StopBadware website due to Cloudfare cache management.

Table 2.7: Number of cleanup website visitors per treatment group.

Treatment type	Campaign 1		Campaign 2	
	Host. Provider	Owner	Host. Provider	Owner
University	4	1	5	3
Indiv. researcher	1	2	3	5

difference among visitors and non-visitors in terms of cleanup rates. The average cleanup time for the hosting providers that visited one of our websites was around 2 days, while for non-visitors it was almost 5 days on average. This decrease in average cleanup time may indicate a positive impact of the cleanup website. To further analyze the impact of this variable on the cleanup process, we estimate the survival probabilities for hosting providers that visited versus those who did not visited the cleanup website (see Figure 2.7). This figure shows that after 3 days, those hosting providers that visited one of the cleanup websites had already cleaned 53.8% of the infected domains, while those who did not visit any of our cleanup websites had only cleaned 28.8% of the infected websites after 3 days. However, though the cleanup rate is quite different during the first 3 days since the notice was sent, the survival curves are not significantly different (Log-rank test: $\chi^2 = 1.5, p = 0.214$). Thus, after the 16-day followup the cleanup rate of the hosting providers that visited our websites is not significantly different from the cleanup rate of those who did not visit our website.

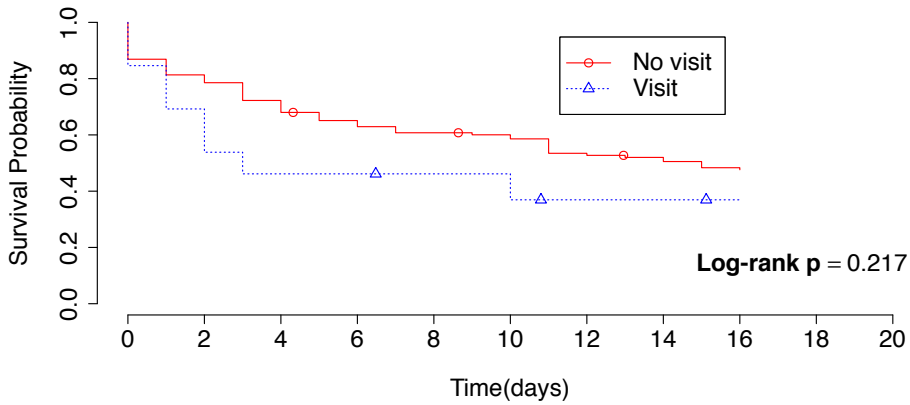


Figure 2.7: Survival probabilities per cleanup website hosting provider visits

This also suggests that hosting providers have different policies to deal with website infections. Table 2.8 describes some basic statistics of the top 10 autonomous systems in terms of number of Asprox infected domains. We can see clear differences both in terms of the amount of remediated infections and also in terms of average time to clean up an infected website. For instance, 'InMotion' hosting provider remediated all the infection in less than 4 days in average, while 'OVH' only remediated 21.05% of the websites and took around 8 days on average for those it did clean up. Figure 2.8 plots the survival curves for these hosting providers. Again, we can see significant different in terms of cleanup rate for the different hosting providers. 'InMotion', 'CS Loxinfo' and 'Hetzner' had cleaned more than 20% of their infected websites after 5 days while the rest of hosting providers took more than 10 days to achieve a similar percentage.

Table 2.8: Summary cleanup statistics per AS owner.

AS Name	#AS	# Infections		% clean		Avg. Cleanup Time (days)		CC
		Camp. 1	Camp. 2	Camp. 1	Camp. 2	Camp. 1	Camp. 2	
CloudFlare	13335	0	9	-	44%	-	10.25	US
OVH	16276	9	29	22.22%	21%	10.00	7.29	FR
InMotion-West	22611	2	6	100.00%	100%	7.00	5.17	US
Hetzner	24940	5	15	100.00%	20%	5.20	1.67	DE
Dreamhost	26347	0	6	-	33%	-	6.50	US
SoftLayer	36351	3	25	66.67%	20%	8.33	4.40	US
SadeceHosting	42910	2	9	50.00%	11%	10.00	7.00	TR
InMotion	54641	0	6	-	100%	-	3.33	US
Strato	6724	1	12	100.00%	25%	10.00	5.40	DE
CS Loxinfo PLC	9891	0	17	-	71%	-	3.08	TH

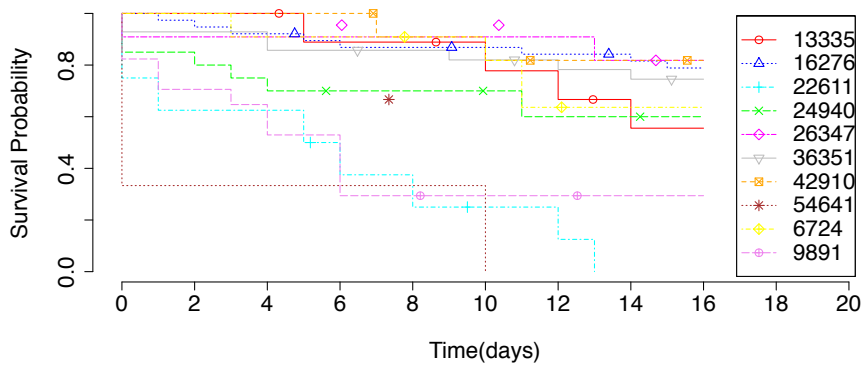


Figure 2.8: Survival probabilities top 10 autonomous systems

Similarly, we measured whether website owners that visited our websites were capable of cleaning their infected websites faster. The average cleanup time for the website owners that visited one of our websites was 4.20 days in average, while for those who did not visit a cleanup website it was 4.26 days in average – an insignificant difference. The same result is shown by the survival probabilities (see Figure 2.9). After 7 days, the owners who visited the site had cleaned 36.4% of the infected domains, while those who did not visit cleaned 40.8% of the websites after 7 days. Thus, visiting the cleanup website did not make a difference for the website owners (Log-rank test: $\chi^2 = 0.2, p = 0.648$). In short, it seems providing cleanup advice is not helpful, at least not in this form. If we assume that less technically competent owners are more likely to follow the link, then even basic advice does not enable them to achieve better cleanup.

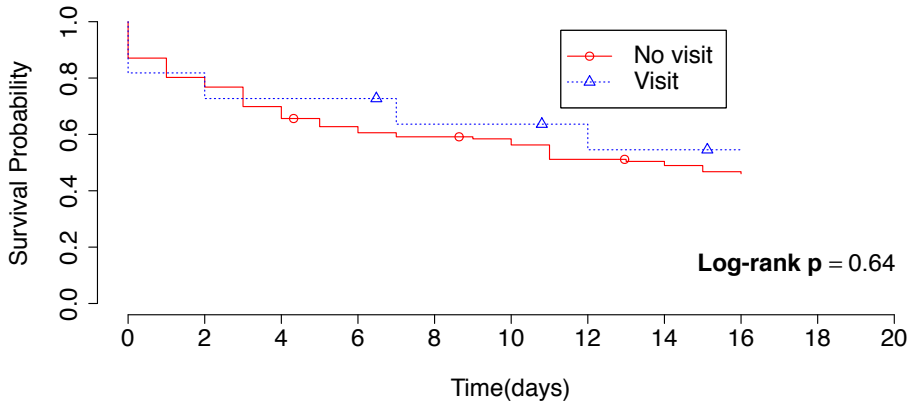


Figure 2.9: Survival probabilities per cleanup website owner visitors

These results suggest that: i) hosting providers play a major role when it comes to remediating an Asprox infection, ii) hosting providers that visited our cleanup website correlated to a higher rate of remediating the infection that those that did not, and iii) website owners seem to not have enough skills to clean up their own website once it gets infected, even when basic suggestions are provided.

2.4.3 Analyzing responses from notified parties

During our experiment, we contacted 480 abuse contacts and received e-mail responses from 89 contacts. Of these 11 (12%) were clearly from a human, while 78 (88%) were machine-generated. The vast majority of responses were in English. Other common languages included Chinese, Russian, German, French, Turkish, Iranian, Thai, and Spanish.

Automated messages came in two forms: confirmations (28%) and tickets (72%). Confirmation e-mails simply acknowledge receiving our notification. Tickets provided a reference or ticket identifier associated with our notification message.

Throughout the experiment, 173 out of 240 notifications we sent to site owners bounced back mostly due to lack of `abuse@domain` address. On the other hand, the same addresses belonging to hosting providers bounced back once, indicating that the vast majority of hosting providers were at least setup to receive abuse e-mails. The difference can be explained in terms of awareness, technical knowledge, and/or liability. Whereas site owners are likely not aware of abuse reporting conventions, lack technical knowledge, and generally are not held liable for the distribution of malicious content, hosting providers as organizations generally are aware, and also potentially liable [59].

Table 2.9: Summary statistics on the cleanup time according to the type of response

Treatment Group	Campaign 1						Campaign 2					
	Human responses			Automated responses			Human responses			Automated responses		
	#	% clean	Median Cleanup	#	% clean	Median Cleanup	#	% clean	Median Cleanup	#	% clean	Median Cleanup
Indiv. Researcher	3	100%	1 day	7	86%	5 days	1	100%	1 day	16	56%	13 days
University	1	100%	2 days	5	60%	12 days	4	75%	5 days	23	57%	4 days
Anti-malware Org.	1	100%	6 days	7	100%	2 days	1	100%	4 days	20	60%	4 days

We investigated the relationship between the responses of notified parties and their cleanup behavior. Table 2.9 provides some summary statistics regarding the status of the infected URLs after 16 days according to each response type that we received. Entries are given for each treatment group. Again, we reported the percentage of websites that have been found clean at the end of our 16-day investigation and the median number of days required to clean up those sites. We cannot observe any significant difference in the number of received responses across the treatment groups. This suggests that none of the notified entities decided whether to reply based on the reputation of the sender.

We did, however, find statistically significant differences between each of the type of responses and cleanup rates (Log-rank test: $\chi^2 = 16.6, p = 0.000247$). As shown in Figure 2.10, within four days after notification, 64% of human responders had already cleaned up their websites, while automated responders had remediated 43% of the infections, and those parties that didn't reply at all had only cleaned 29% of the compromised sites. Thus, the second strongest reactions came from contacts configured to send automated responses. This indicates that hosting providers using a system to automatically process notifications and complaints are more likely to act. As expected, the least effective reaction came from those hosting providers that never responded. After the first week, only 32% of such contacts had conducted some remediation; after 16 days, 48% had. While these cleanup rates are lower, they do show that even when hosting providers do not respond, it does not imply

they ignored the message.

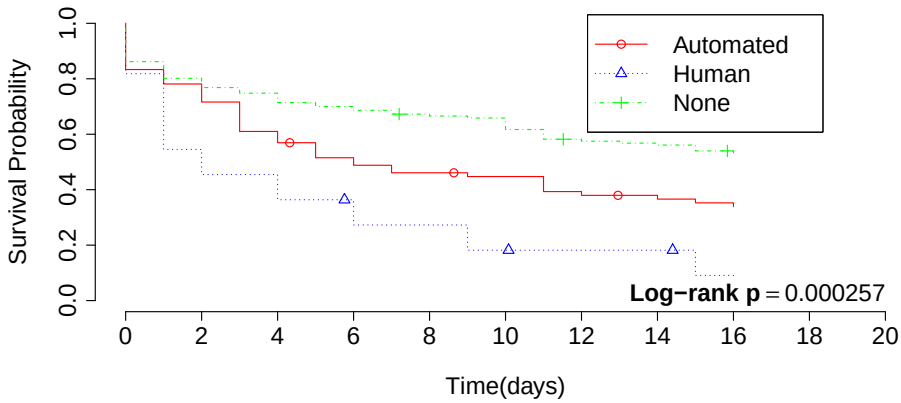


Figure 2.10: Survival probabilities per response type

2.5 Related Work

A few researchers have recently begun investigating how notifications about system compromise or vulnerability can promote remediation. Most similar to our own work, [32] conducted an experimental study on web-based malware URLs submitted to the StopBadware community feed. They found that abuse reports sent with detailed information on the compromise are cleaned up better than those not receiving a notice (62% vs. 45% cleaned after 16 days). Moreover, they found no difference between the cleanup rates for websites receiving a minimal notice and those not receiving any notice at all. Based on this finding, we elected to provide detailed information in the abuse reports we sent. Thus, we corroborate their finding that detailed notices work on a different type of incident dataset.

Furthermore, we studied how different forms of notifications affected uptimes of malware cleanup rates [36]. To this end, we compared the uptimes of ZeuS command and control servers provided by Zeus Tracker, Cybercrime Tracker and a private company. ZeuS Tracker and Cybercrime Tracker present a publicly accessible dynamic webpage that displays ZeuS malware command and control servers. On the other hand, the private company did not publicize any of detected command and control servers. We showed that publicized command and control servers were mitigated 2.8 times faster than the ones that were not publicized.

Another malware-orientated study supported the notion that notifications spur intermediaries to take action: in [34], researchers setup vulnerable webservers and compro-

mised them. After a period of 25 days, they notified their own web hosts. Approximately 50% took action, generally suspending access. To ensure that the notifications were actually being read and not simply being acted upon without evidence, false abuse reports were also sent, resulting in 3 of the 22 providers suspending an account without actual evidence. This in turn suggests that most, but not all, recipients investigate abuse reports before taking action.

Whereas the present work and studies described above focus on reports of compromise, other researchers have sent notifications to the operators of vulnerable, but not necessarily compromised, systems. The goal here is to patch the vulnerable systems instead of remediating an infection. For example, [41] notified hosts vulnerable to the widely reported Heartbleed vulnerability. After scanning and excluding device and large-scale cloud providers (such as Amazon), researchers automatically identified 4,648 unique administrative contacts for 162,805 vulnerable hosts. They then divided the contacts into a treatment group receiving notifications and a control group that did not (at least initially). The treatment group was notified by e-mail and pointed to a detailed patching guide hosted at a University website. The researchers observed a 39.5% rate of patching for those receiving notifications, versus 26.8% for those that did not.

Similarly, [42] issued notifications for systems vulnerable to DDoS amplification attacks involving NTP. Rather than directly notify each individual host with information about the vulnerability, the researchers provided lists of afflicted IP addresses to key organizations such as abuse team contacts at CERTs, security data clearinghouses such as Shadowserver, and afflicted vendors such as Cisco. They complemented this effort by working with CERTs to issue informative advisories warning of the vulnerability and how to patch affected systems. This multi-pronged approach proved very effective: they observed a 92% reduction in amplifiers after three months tracking a population of 1.6 million affected hosts. Although the authors did not design an experiment with a control group, the researchers credited the campaign's success to collaboration with reputable sources who then issue notifications. This suggests that sender reputation might be influential after all, despite the negative findings from our study. In future work, we recommend investigating alternative sources of reputation, such as other intermediaries capable of coordinating cleanup and/or the use of private contact details for sharing compromise information.

Finally, with respect to general e-mail spam, a quasi-experiment by [37] saw researchers use two blocklists to compile a large source of e-mail spam and publish aggregated measures on SpamRankings.net. They then published the results for a treatment group and withheld results for a control group, observing a 15.9% reduction in spam among the treated group. Rather than notify individual hosts in order to remediate infections, the researchers' strategy relied on public shaming. The study indicates that abuse information could provide incentive for intermediaries to cooperate in remediating abuse on their networks.

2.6 Limitations

A number of limitations may impact the findings from our study.

First, we selected contacts to notify by inspecting the WHOIS for affected domains. Many abuse reports are sent between personal contacts, not general contact addresses, but we were unable to capture the impact of reputation in these trusted interactions. Our findings, therefore, apply only to the baseline case where personal contact has not been established. To put it differently, we are not claiming that reputation does not matter. Not only did an earlier study suggest it might (see section2.5), but the actual practices of abuse reporting show this every day. For example, many providers work with trusted reporters. In some cases, these notifications are trusted enough to allow for automated countermeasures or takedown actions.

Second, we measured reputation by the domain associated with the notification and the website used for cleanup advice. One potential issue is that our University-affiliated address was `tudelft.nl`, as opposed to the more widely known `.edu` top-level domain.[2] Nonetheless, anyone visiting the website for cleanup advice would clearly see the association with a University, while those visiting StopBadware's website would see that it was a non-profit cybersecurity organization. However, this is only one way to measure reputation. Reputation can also be established by sending credible notifications over a period of time. Because none of the organizations in our study regularly send notifications, we were unable to measure reputation in this fashion. However, it is something that we hope to do in future work, provided that we can partner with an organization that regularly sends abuse reports.

Third, we relied on a source of compromised URLs focused specifically on the malware delivery component of a single, long-established botnet. We made this design decision intentionally, in order to control for the natural variation that exists between different types of abuse data. For example, a hosting provider might prioritize cleanup of command and control infrastructure over hacked websites that deliver malware. Furthermore, advanced persistent threats, banking Trojans and phishing sites could attract more attention from hosting providers due to the financial implications and potential liability. The impact of sender reputation may differ in these scenarios, and so we defer such investigations to future work.

Fourth, there is a chance that latent characteristics appeared disproportionately in the treatment groups that influenced the overall outcome. For example, hosting provider size and type (shared vs. dedicated) may influence cleanup rates, but we were unable to verify that the distribution of these features is proportionate among treatment groups.

Fifth, we did not study re-infection of previously cleaned websites. Frequently, websites are recompromised when the hole that let the attacker in the first time is not plugged [60].

[2]Moreover, in certain cases, e-mails from `.nl` and `.org` addresses get caught in spam filters, whereas those from Gmail get through.

Because we were primarily interested in measuring the response to abuse reports, we elected to ignore subsequent reinfections.

Finally, there are a number of characteristics closely related to reputation that we did not examine. For example, none of our reports carried any suggestions that punitive action may result for ignoring the report. By contrast, notifications sent by Google (who controls search results) or ISPs and hosting providers (who control Internet access) might carry more weight due to the implication that there could be consequences for inaction. We defer investigating these effects to future work.

2.7 Conclusion

In this paper, we described an experiment to measure the differences in cleanup among notifications from senders with differing reputations. We find no evidence that reputation, as measured by the sender's type of organization, influences cleanup rates. However, we do find that detailed notices results in better cleanup overall. This confirms earlier findings carried out on websites distributing drive-by-downloads by [32].

Furthermore, we find that publicizing and linking to a cleanup website containing specific instructions improves the cleanup rate when hosting providers view the instructions. However, this same positive impact is not shared by resource owners who served as point of contact for their domains. This suggests that differences in technical proficiency influence the success of a notification. Finally, throughout the trial, reports that elicited personal responses from the affected parties achieved higher cleanup rates. This suggests that personal interaction may contribute to better cleanup.

The role of the attacker in evading detection also plays a big role in how effective cleanup can be. We presented evidence that when compromise could be easily verified, cleanup rates were much higher than when the attackers took steps to hide the compromise. We plan to study this effect in greater detail in future work.

Moving forward, we recommend three specific areas of study to further build on the work of this paper: first, the content of the notification and the presence of punitive measures; second, studying how cleanup websites are actually used by resource owners and intermediaries in order to craft a more effective message; and finally, sending notifications for other aspects of the cybercrime ecosystem, including command and control.

CHAPTER 3

Measuring the impact of large-scale vulnerability notifications

In the previous chapter, we measured the effectiveness of abuse notifications. In this chapter, we looked into the feasibility and effectiveness of large-scale vulnerability notifications to intermediaries and resource owners. As large-scale vulnerability detection becomes more feasible, it also increases the urgency to find effective large-scale notification mechanisms to inform the affected parties. Researchers, security companies and other organizations with vulnerability data have a variety of options to identify, contact and communicate with the actors responsible for the affected system or service. A lot of things can – and do – go wrong. It might be impossible to identify the appropriate recipient of the notification, the message might not be trusted by the recipient, it might be overlooked or ignored or misunderstood. Such problems multiply as the volume of notifications increases. In this chapter, we undertake several large-scale notification campaigns for vulnerable servers. We investigate three issues: What is the most effective way to reach the affected parties? What communication path mobilizes the strongest incentive for remediation? And finally, what is the impact of providing recipients a mechanism to actively demonstrate the vulnerability for their system, rather than sending them the standard static notification message.

3.1 Introduction

The Internet's decentralized and trans-boundary architecture requires effective voluntary collaboration between defenders to fight off security threats. This can take the form of abuse reporting, where one party notifies another of an abuse incident and asks it to act against the abuse. Another important collaborative mechanism is to detect and remediate vulnerabilities before they are exploited by notifying the entity responsible for the vulnerable system or service.

Notifications that drive such voluntary cyber-defense take on many forms, from manually crafted emails sent to webmasters all the way to machine-generated feeds that recipients can tailor to their information needs. Some notifications are unsolicited and pushed to

recipients, others require the recipients to take action and request data via APIs or mechanisms. Despite differences in the content, context and technology of how the countermeasures are deployed, each is premised on some type of notification about the abuse or vulnerability, being sent from one party to another.

In this paper, we focus on vulnerability notifications. They have been around for quite a while. The security community, however, has only recently started to study the effectiveness of these mechanisms. We know remarkably little about the aspects and factors that drive higher vulnerability remediation rates and how recipients feel about various types of notifications [61, 43]. Moreover, there is a lack of evidence-based guidelines on how to make large-scale notification mechanisms more useful and effective in remediating vulnerabilities.

Any large-scale notification mechanism will have to decide on a variety of issues regarding how to get the vulnerability information in the right hands and how to incentivize actual remediation. In this paper, we investigate three issues: What is the most effective way to reach the affected parties at scale? What communication path mobilizes the strongest incentive for remediation; contacting the nameserver operator directly, their customer or the network operator? And finally, what is the impact of providing recipients a mechanism to actively demonstrate the vulnerability for their own system, rather than sending them the standard static notification message. We study these questions by undertaking several large-scale notification campaigns for authoritative nameservers that are vulnerable to so-called ”zone poisoning” [62].

In the next section, we outline the methodology used for this experiment. The results of the experiment are explained in Section 3.3. In Section 3.4, we present an explanatory analysis of email bounces and remediation. We explore reactions of email recipients in section 3.5. Finally, we compare our findings to the related work in section 3.6 and we summarize our conclusions in section 3.7.

3.2 Methodology

We designed an experiment around nameservers that are configured to allow non-secure dynamic updates. This allows for an attack called *zone poisoning*. In this section, we explain the overall design of the study, which is summarized in Figure 3.1. First, we briefly describe the vulnerability and how we identify vulnerable nameservers. Then we outline the three notification campaigns using different communication channels: nameserver operators, domain owners and network operators. Subsequently, we discuss the experimental design that was used in each campaign to test the impact of the different notifications. We describe content of the notifications, the demonstration website and the recipient survey. Fourth, we describe our rationale for constructing the experimental groups. Fifth, we discuss the ethical issues associated with our approach. Finally, we explain how we evaluate the results.

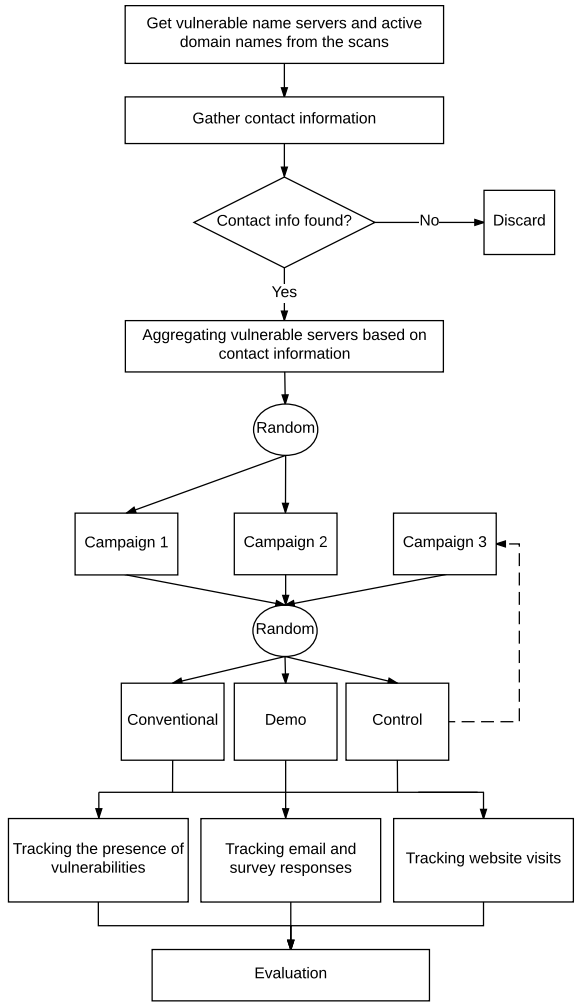


Figure 3.1: Flow diagram of the progress through the phases of our experiment

3.2.1 Vulnerability: Non-secure DNS Dynamic Updates

Korczyński et al. presented a measurement study of authoritative nameservers that allow non-secure dynamic updates [62]. This vulnerable configuration allows anyone on the

Internet to freely manipulate DNS entries in the zone files of that authoritative nameserver. This attack is referred to as *zone poisoning*. The attack is as simple as sending a single DNS dynamic update packet that is compliant with RFC 2136 [63] to a non-secure server. In the simplest version of an attack, a miscreant could replace an existing `A` or `MX` resource record in a zone file of an authoritative server and point the domain name to an IP address under control of an attacker. This can be used, for example, for phishing or for intercepting email by changing the record for `mail.domain.com`. The requirements for the attack to succeed are: non-secure updates are allowed by an authoritative server for a given zone and the miscreant knows the domain name and its nameserver. Finding this information is trivially easy. In short: it is a serious security threat. The original paper [62] discusses the threat model in more detail.

Korczyński et al. analyzed a random sample of 2.9 million domains and the Alexa top 1 million domains and found that at least 1,877 (0.065%) and 587 (0.062%) of domains are vulnerable, respectively. Among the vulnerable domains were governments, universities and banks, demonstrating that the threat impacts important services.

The first measurement study was extended from a sample to a comprehensive scan of the domain name space. Between September 21 and October 11, 2016, Korczyński et al. performed a global scan of the non-secure DNS dynamic updates and found 309,687 vulnerable domains and 5,738 IP addresses with vulnerable authoritative nameservers. In total, they counted 579,186 unique "domain, nameserver" tuples. Here, we limited our study to the 21,506 domains that were active during the period of our experiments – which corresponds to 4,149 IP addresses of the vulnerable nameservers with Start of Authority (SOA) records.

3.2.2 Experiment

In this section, we outline the research questions which we attempted to answer via the experiment. We were specifically interested to answer following research questions:

1) How Can You Reach Resource Owners at Scale?

Security researchers have a variety of options to identify the contact details of owner, operator or user of the vulnerable resource. One approach is to use dedicated mail aliases as mentioned in the RFC 2142 for abuse and network-related problems [50]. For DNS-related problems, the RFC says to use the SOA `RNAME` field to provide contact information for the zone's administrator. Moreover, the RFC defines "hostmaster" as the mail alias to be used for DNS issue. It also mentions "abuse" as the email aliases that can be used for generic abuse and vulnerability notifications.

During the first campaign, we test the effectiveness of reaching administrators of vulnerable nameservers by sending a notification to the email as specified in the SOA `RNAME` field. When this field was not present, we used the "abuse" email alias.

During the second campaign, notifications were sent to the owners of vulnerable domains. We obtained the contact details from the registrant's email address in the WHOIS

records of the domain. When we couldn't find the registrant's email address, we sent the notification to <hostmaster@domain>. Furthermore, the "abuse" email alias for domain was used as a fallback option when a bounce report of the initial notification was received.

2) Which Channel Contains the Strongest Incentive for Remediation?

Next to getting the notification to the chosen recipient, there is also the issue of whether that recipient has an incentive to perform remediation. Since there are different affected parties that could be notified of this vulnerability, we wanted to test whether it was more effective to contact resource owners directly, to go via their customers or to go via their network operators. The direct route seems the most obvious communication channel, but the name server operator might not have an incentive to remediate. The domains threatened by zone poisoning might not be his. Changing the configuration to a secure mechanism for dynamic updates might also generate cost, for example, to replace this functionality with what is inevitable a more complicated solution than the non-secure configuration. Under these conditions, it might be rational to wait and see whether actual abuse will occur and with what frequency.

The domain owners, which are typically the customers of the nameserver operator, might care more about protecting their domain. Our notification suggested that they might have to contact the nameserver operator, for example their hosting provider, to ask for the problem to be remediated. The operator probably has a stronger incentive to act on such a customer request than on the friendly advice of an academic research team. We tested which path leads to better remediation by contacting different recipients in each of the three campaigns. First, we notified nameserver operators directly via SOA `RNAME` field. Second, we contacted domain owners via the registrant's email address in the domain `WHOIS` record. Third, we would notify the next higher level intermediary, the network operator, via `IP WHOIS` abuse contact field.

3) Does a Demonstration of the Vulnerability Produce Better Remediation?

Since recipients might receive many vulnerability notifications, something that will only increase with the rise of large-scale vulnerability detection, they are probably not willing or able to act on all of them. It seems inevitable that recipients some form of triage the incoming messages, if only in the form of ignoring those that do not seem trustworthy, credible or critical.

Providing recipients with a simple way to demonstrate the vulnerability for their own nameserver or domain, would allow them to immediately verify the trustworthiness, credibility, and criticality of the notification. To test whether this improved remediation, we built a website that demonstrated the vulnerability. Recipients could let the site inject a harmless record in the zone file of a vulnerable domain. The site would show the existence of this new DNS record, proving that anyone on the Internet could change any DNS record for that domain. (We included controls to avoid abuse, recipients could only test their own domains or nameservers.)

To test whether the demo makes a measurable difference, we designed two different treatments: one standard notification message and one notification message that included

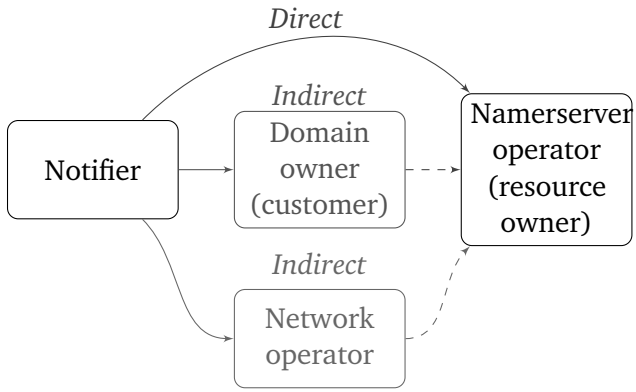


Figure 3.2: Communication channels per campaign

the same information plus a link to a website that we built that demonstrated the vulnerability. In each of the three campaigns, the recipients were assigned to one of three groups: a control group that received no notification; a treatment group receiving a conventional notification; and a second treatment group receiving a notification message with a link to a site we built where the vulnerability is demonstrated. We discuss the notification content and website in more detail in subsection 3.2.4.

3.2.3 Group Assignment

There are several steps in the overall experiment, as illustrated in Figure 3.1. It is a bit complicated, but the easiest way to think about is this: the experiment with the two notification treatments (notification with and without access to the demonstration website) is repeated three times, once for each communication channel (see Figure 3.2).

We chose for three sequential, rather than parallel, campaigns to keep the experiment manageable and to prevent possible contamination. Sequential campaigns could have cause contamination in various ways. For example, if the first 2 campaigns would have run sequentially, once we contacted a domain owner, then she might have contacted the nameserver operator, as we hope she would. The operator, however, might be responsible for other nameservers or domains as well, which might be in another treatment group or in the control group in the other campaign. As a result of this, same nameserver operator will appear in different treatment groups, thus receiving different treatments.

There are several assignment processes during the study. The process starts with identifying the relevant contact point. For each vulnerable nameserver, we extracted the email address of the person or organization responsible for the DNS zone from the correspond-

ing SOA record. In 256 (out of 4518) cases, the SOA record for the nameservers was not present, hence we removed the nameserver and associated domain names from our study. Next, we aggregated the nameservers and domain names by unique SOA contact information. This resulted in 3967 unique nameserver contacts. We then randomly assigned each contact to the first or second campaign (see Figure 3.2).

For the first campaign, the nameserver operators were randomly assigned to one of three groups: control, conventional notification and demonstrative link. The contacts assigned to the control group received no notification during this campaign. As we discuss below, we did notify them later on in the study. The measurement period of the first campaign lasted 19 days. We tracked remediation, survey and email responses and website visits.

Once the first campaign was done, we moved to the second campaign. First, we checked whether the domains and nameservers assigned to this campaign were still vulnerable and found that 70 (out of 1984) cases were remediated without being notified by us. And we also checked whether units assigned to second campaign shared any IP address or domain with the previous campaign and found that 451 (out of 1914) cases were sharing at least one IP address or domain. They were removed from the experiment. To identify the relevant contact information, we had purchased `WHOIS` data[1] and extracted the registrant's email address. We did not use any other email address field in the `WHOIS` record, as they could lead to the hosting provider or another entity. When registrant's email was missing, we generated an email address using the "hostmaster" email alias, as recommended by RFC 2142 [50].

Next, we conducted one more aggregation. If two nameservers had different names but they both resolved to the same IP address, then we bundled them, and the associated domains, together. This was done to further reduce the risk of contamination. We then randomly assigned each unique nameserver contact point (or bundle thereof) to one of three groups: control, conventional notification and notification with link to demonstration website. All domains associated with a nameserver contact would receive the corresponding treatment assigned to that contact. For example, for all domains that ended up in the conventional notification treatment group, we contacted the registrants with the conventional notification message. It is important to reiterate that in the second campaign we did *not* contact any nameserver operator directly.

Once the measurement period of the second campaign ended, we took the control groups of the first and second campaign as the subjects for the third campaign. First, we checked that the domains and nameservers were still vulnerable. As it turns out that 95% of these hosts were still vulnerable. We extracted contact information for the network operators by querying Abusix's Abuse Contact DB [64] for the `IP WHOIS` abuse contact that belongs to the IP address of the vulnerable nameserver. These abuse contacts belong to upstream intermediaries, such as ISPs and hosting companies. Next, we aggregated the vulnerable domains and nameservers per unique abuse contact point. We then randomly

[1] We purchased `WHOIS` domain data from `whoisxmlapi.com`.

assigned these contacts to one of the three treatment groups, as was done in the first and second campaign.

3.2.4 Notifications, Demonstration Website and Survey

Notifications for both treatments were sent from the same dedicated email account belonging to Delft University of Technology. To reduce the risks of unsuccessful email transmission, we disabled inbound and outbound spam filters used by the university.

The conventional notification treatment consisted of an email with a plain text vulnerability report. It contained a brief explanation of how we discovered the vulnerability, what the security impact is if it is abused, and how it can be remediated. We enumerated the vulnerable nameservers or domains associated with the contact point. The message concluded with a link to a short survey. The other treatment consisted of basically the same notification, plus a link to the vulnerability demonstration website. Full details of the notification messages can be found in Appendix B.

We built and operated the demonstration website. Figure B.5 in appendices shows screenshots of the interface. The site provided recipients with an opt-in tool that would provide a live demonstration of the vulnerability for their nameserver or domain – that is, an actual record, albeit a harmless one, would be injected into the zone file. The new record added a subdomain called *zonepoisoning* to the vulnerable domain. This sub-domain would then correctly resolve through DNS and point to a webserver belonging to our experiment, showing that the record was successfully inserted. The added record remained in the zone file for 10 minutes, after which it was removed automatically. After every interaction, website interface shows the results of the subdomain injection attempt. Vulnerable servers trigger an interface where a link to created subdomain and an explanation is displayed to verify the existence of the vulnerability. On the other hand, patched servers triggered a different interface, explaining the unsuccessful injection attempt.

The website, the server to which the new subdomain resolved and the server used for the scans for vulnerable nameservers and domains all provided information on how to opt out of our study. To prevent potential abuse, we provided recipients with a link containing a unique token that allowed us to restrict what domains or nameservers could be tested by the visitor of the website. Recipients could only demonstrate the vulnerability for the nameserver, domain or networks for which they were the contact point.

The website and the notifications included a link to a short survey where the recipients were asked to answer several questions about our notification process. The questionnaire was designed to capture the recipients' reaction to our notifications, to notifications in general and to the way we conducted our research.

3.2.5 Tracking Process

To track remediation during each campaign, and to update our data on vulnerable nameservers and domains, we performed 7 scans between November 3 and December 29, 2016.

We used the scanner that was developed by Korczyński et al. [62]. It sends a DNS update request packet that is compliant with RFC 2136 [63]. The request was to add an extra `A` record to the zone file, associating a new subdomain (e.g., `researchdelft.example.com`) with the IP address of the web server of our project. When a nameserver operator would visit the IP address, she would encounter a page with an explanation of the study and an easy opt-out mechanism (see Figure B.3 in Appendix B).

Our scanning setup was designed to have minimal impact, while also taking into account random packet losses. We first sent two DNS update request packets. We then performed four DNS lookups, from two different measurement servers, to verify if the added domain correctly resolved to our web server's IP address. Next, we removed the test DNS record by sending a delete update request. Finally, we queried the authoritative DNS server and try to resolve the subdomain once more, in order to confirm that the added record was successfully deleted.

We considered an authoritative nameserver as remediated if it no longer appeared vulnerable in any subsequent scan. A domain was considered as remediated if none of its authoritative nameservers are found to be vulnerable.

3.2.6 Ethical Considerations

Our study aims at improving the deliverability of vulnerability information to owners of computing resources, such as websites or servers. Vulnerability notifications are a well-established practice to help operators of vulnerable resources to better protect themselves against criminals who might abuse the vulnerabilities.

The only valid method available to detect and demonstrate the vulnerability was to insert a benign record into a zone file. We weighed the tradeoffs and decided that the benefit of helping the server operators to protect themselves outweighed the potentially intrusive nature of the scans. The ethical considerations are discussed in more detail in [62]. The inserted records were only present for a very short time. We did not interact with any of the existing records in the zone. We did not observe or hear about any problems with the vulnerable servers because of our scans, as we expected, since our interaction with the servers was fully compliant with the relevant standard. Furthermore, recipients were provided with an opt-out mechanism in every engagement. During the study period, only one recipient asked to be excluded of the study.

3.2.7 Evaluation

To assess which communication channel contains the strongest incentives for remediation, we evaluate the results based on two metrics: (i) reachability, i.e., the email bounce rate;

and (ii) the remediation rate. We measured the impact of the vulnerability demonstration by comparing the remediation rate of the recipients who visited and/or used the demonstration tool versus those who did not. In addition, we explored the email and survey responses to learn more about how various recipients perceived our vulnerability demonstration website and the content of the notifications.

3.3 Notification Results

In the previous section, we outlined the experimental design, methodology and objectives. In this section, we present the results of each campaign on the deliverability of notifications and on the remediation rate. Next, we discuss the efficacy of the demonstration website. We end with a comparative analysis of the communication channels.

3.3.1 Notification Deliverability

In this section, we analyze the deliverability rates of the notifications. Table 3.1 summarizes the bounce rates per campaign.

Table 3.1: Bounce rates

Campaign	Treatment type	Total number of aggregated contacts	Number of emails initially send	Rate of undelivered emails	Number of fallback emails send	Rate of undelivered emails
1	Demonstration	669	669	70.40%	357	82.07%
	Conventional	657	657	67.73%	335	86.26%
2	Demonstration	451	940	44.68%	279	88.88%
	Conventional	451	1111	35.64%	282	89.00%
3	Demonstration	184	208	12.01%	–	–
	Conventional	183	209	5.2%	–	–

First Campaign

Reaching the relevant contact points at scale turned out to be a huge problem. As shown in table 3.1, initially 669 emails were sent with a link to demonstration website and 657 emails with conventional content. Of these 669 emails for the demonstration group, 70% returned a delivery failure. Similarly, 67.73% of the emails with conventional notifications failed to be delivered.

To reach more affected parties, we sent a second email when the first one had generated a failure. This second email was sent to an address we generated in compliance with RFC 2142 [50], of the form `<abuse@domain.com>`, where the domain corresponded

to the nameserver domain. So the operator of `ns1.example.com` would be contacted at `<abuse@example.com>`. We sent an additional 692 emails this way: 335 for the conventional treatment group and 357 for the demonstration group. This second attempt incurred an even higher bounce rate: on average, 84% of these messages generated a delivery failure.

Second Campaign

We sent 2,051 emails to domain owners in the second campaign: 1,111 emails to the conventional notification group and 940 emails to the demonstration group. Of these 2,051 emails, 39.78% bounced on average. The rate was slightly higher for the demonstration group. Similar to the first campaign, when a notification could not be delivered, we applied a fallback option. We sent a second email to `<abuse@domain.com>` addresses for vulnerable domains. In total 561 emails were sent in hope to reach more vulnerable domain owners. Around 89% of these bounced also.

Third Campaign

We sent 417 emails during the third campaign. For network operators, as identified via the `IP WHOIS` abuse contact for the IP address of the vulnerable nameserver, reachability was much better. Only 36 out of the 417 notifications generated a delivery failure. For this reason, we did not use a fallback option.

3.3.2 Remediation Rates

The reachability of nameserver operators was poor, for domain owners it was slightly better and for network operators it was quite good. This raises the question of whether this is also connected to a difference in remediation. In this section, we analyze the remediation rates of the different treatment and control groups for each campaign.

Table 3.2: Summary statistics remediation per treatment group, counted per unique SOA contact points

Treatment Type	Campaign 1				Campaign 2				Campaign 3		
	#	After 3 days	After 13 days	After 19 days	#	After 3 days	After 13 days	After 19 days	#	After 3 days	After 13 days
Control	657	3.04%	4.26%	5.02%	476	2.31%	3.78%	4.62%	320	0.3%	1.87%
Demonstration	267	12.35%	14.23%	18.35%	345	5.50%	6.37%	10.14%	382	4.97%	8.11%
Conventional	260	8.84%	9.61%	14.23%	327	5.81%	6.72%	12.23%	329	3.03%	5.77%

Table 3.2 provides a summary of the status of the vulnerable servers during three different measurements in the first and second campaign, and during two measurements for

the third campaign. The additional third measurement for the first two campaigns allows us to see the impact of the fall-back notifications. The table reports the percentage of contact points that took action, excluding those that we could not reach. Overall, remediation rates were low. The highest rate for any group or campaign was 18% of all vulnerable nameservers.

First Campaign

Notification clearly makes a difference. In the control group, 5% of the contact points remediated the vulnerability within 19 days, compared to 18% and 14% for the two treatment groups. We did a log-rank test and found that difference between the treatment groups and the control group is significant ($\chi^2 = 41.1$, $p = 1.44e - 10$), while the difference between the two treatments is not ($\chi^2 = 1.8$, $p = 0.182$). In short, the demonstration did not make a difference.

Second Campaign

The pattern for the second campaign is similar: notifications increase remediation, compared to the control group (log-rank test: $\chi^2 = 41.1$,$p = 1.44e - 10$), but there is no significant difference among two treatments ($\chi^2 = 1.8$, $p = 0.182$). The remediation rates turned out to be slightly lower when contacting nameserver operators via their customers, compared to the first campaign, where we contacted them directly: 11% versus 16%, on average.

Third Campaign

Since the third campaign focused on abuse contacts at network operators, we aggregated the vulnerable nameservers per network operator. Table 3.3 mentions the remediation rate per recipient. Note that these numbers are different from Table 3.2, as the latter standardized all rates on unique SOA contact points, to make the number comparable. The pattern is basically the same as for the first two campaigns. After 13 days, 15% of the demonstration and and 8% conventional notification groups achieved respectively. Again, a log-rank test concluded that control and treatments were significantly different, while the treatment groups were not.

There are two key findings from these remediation rates. First, providing a vulnerability demonstration to recipients had no observable impact on remediation for any of the contacted parties. Second, there is a modest, but significant difference between the direct and indirect communication channels by comparing the percentage of contact points that took action (see table 3.2). Figure 3.3 plots the survival probabilities. The remediation

Table 3.3: Percentage of remediation by network operators in third campaign

Treatment Type	#	After 3 days	After 13 days
Control	183	0.54%	3.27%
Demonstration	164	9.75%	14.63%
Conventional	173	5.78%	8.09%

rate of the first campaign, which contacted the nameserver operator directly, was slightly higher than during the two indirect campaigns (log-rank test:$\chi^2 = 5.2$, $p = 0.022$).

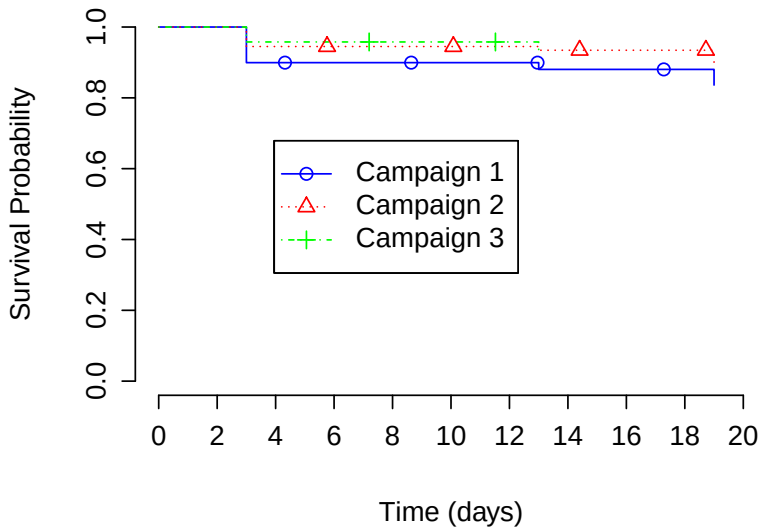


Figure 3.3: Survival probabilities across the campaigns

3.3.3 Efficacy of the Demonstration Website

As a part of our experiment, we built a website that could be used by recipients to demonstrate the vulnerability for their own nameserver or domain. We had two slightly different versions for domain owners and nameserver operators, so that we could tweak the language to their situation. The version for nameserver operators also contained more technical information to assist with the remediation process.

During the experiment period, we tracked the visitors of both websites. As it turns out, most recipients of the link did not visit the website. The number of visitors is presented in Table 3.4. In the first campaign, only 12.2% of the operators visited the website. Those that did made 192 injection attempts. Only about half of these attempts were successful in adding a record. The rest of the attempts failed because the visitor tried to inject a domain name that was not associated with their vulnerable nameserver. In the second campaign, only 7.07% of the domain owners who received the link visited the website. Visitors used the demonstration website 81 times in total. Unlike the previous campaign, 82.71% of the injection attempts were successful. The third campaign showed a similar picture: only 14.75% of the recipients visited the site, they made 137 attempts of which 64.23% were successful.

We have no good explanation for why visitors failed so often to demonstrate the vulnerability. To some extent, this is probably trial and error driven by curiosity. Some of the failed attempts, however, reveal usability problems. While we thought we had designed a very simple interface with straightforward instructions, user behavior told us otherwise. The nameserver operators often tried to test nameserver names, rather than the domains of which the zone file was vulnerable. This happened even though the site instructed otherwise, and we supplied them with a full list of domains to test in the notification email and even proposed a specific domain to test in the main part of the text. All in all, this is a painful lesson that it is very easy to underestimate how hard the problem is of usability of user engagement in the area of security. We can add this lesson to the growing body of work in this area [65].

Table 3.4: Summary statistics on demo website visits

	Campaign 1	Campaign 2	Campaign 3
Number of visitors	32	39	27
Number of attempts	192	81	137
Number of successful attempts	104	67	88
Number of failed attempts	88	14	49

To analyze whether the website helped visitors to expedite remediation, we compared remediation rates of visitors and non-visitors. Figure 3.4, 3.5 and 3.6 show the survival probabilities for all campaigns, respectively. Figure 3.4 shows that after 3 days more than 40% of visitors had taken action, while those who did not visit had remediated less than 10%. After 19 days, almost 60% of the visitors took action, while the non-visitors still hovered around 10%. The same pattern emerged during subsequent campaigns. Log-rank tests show that these differences are significant. We have no hard evidence on what caused

the higher remediation rate. The site may have helped, but it is more likely the effect of self-selection. The recipients that were interested in the demonstration website were probably already more willing to act upon the notification.

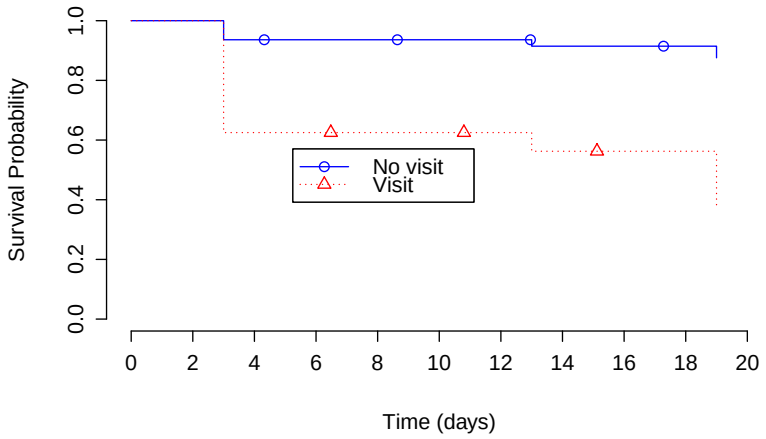


Figure 3.4: Survival probabilities for demonstration website visitors vs non-visitors (Campaign 1)

3.4 Explanatory analysis

We wanted to get a bit more insight into two of the findings of our experiment: the many delivery failures in contacting affected parties and the low remediation rate. For each issue, we discuss several factors and then feed them into a multivariate logistic regression model to analyze their impact.

3.4.1 Modeling Notification Bounce Occurrence

Over the study as a whole, we sent out 5,051 email notifications. Of these, 2,819 triggered delivery failures, a 55.81% bounce rate. We wanted to see if we could explain the probability of a bounce from the features of the recipient's email addresses. We created variables to capture these features.

- *Email Source*: This categorical variable captures the method by which the recipient's email address was obtained. It takes four different values:

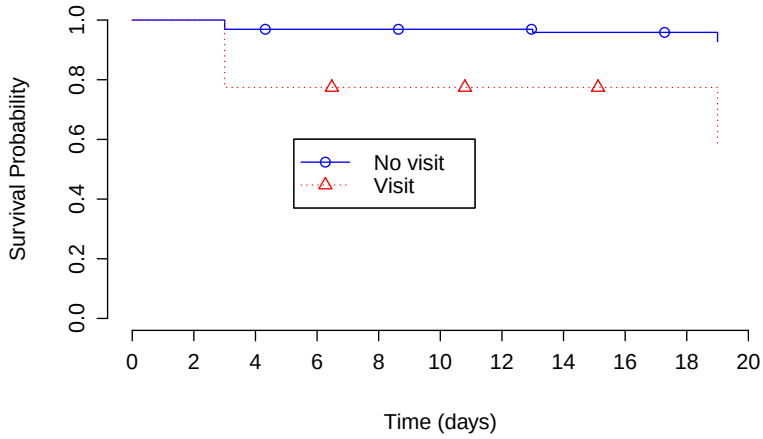


Figure 3.5: Survival probabilities for demonstration website visitors vs non-visitors (Campaign 2)

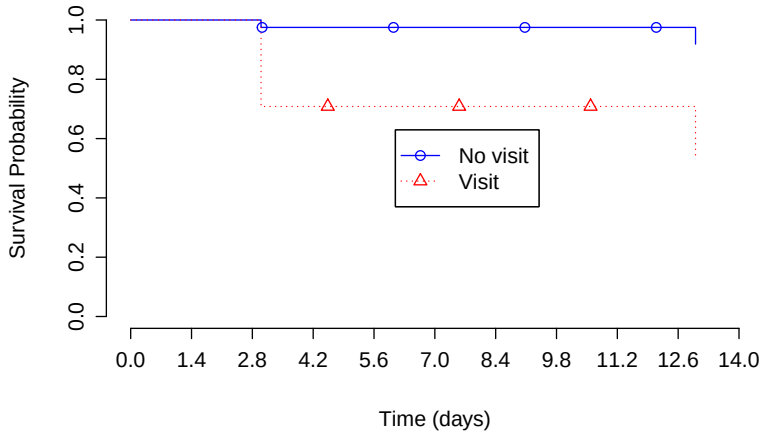


Figure 3.6: Survival probabilities for demonstration website visitors vs non-visitors (Campaign 3)

– x_1: **SOA** : This value represents those notification recipients whose email addresses was obtained by digging the SOA record of the vulnerable nameserver and then extracting the `RNAME` field which contains the email addresses of re-

source owners.

  - x_2: **Domain WHOIS**: This value was set to `TRUE` when the email address was obtained by querying the appropriate `WHOIS` databases corresponding to the gTLD and ccTLD of the vulnerable domains. We then obtained the domain `WHOIS` registrant email field to reach domain owners.
  - x_3: **IP WHOIS**: This value corresponds to those notification recipients whose email addresses were obtained by querying the Regional and National Internet Registry's `WHOIS` databases. We gathered contact details of the entities managing the IP addresses of the vulnerable nameservers.
  - x_4: **Self-generated**: When no contact information was obtained using the aforementioned sources or given information is inacccurete, we generated a RFC-compliant email (i.e., `<abuse@domain>` or `<hostmaster@domain>`).

- x_5: **Privacy-protected Email**: This binary variable is set to `TRUE` when the email address in the `WHOIS` record is behind a proxy service. WHOIS privacy and proxy services are organizations that wish to keep certain information from being made public via WHOIS records [66]. These services can be offered by registrars or their affiliates and they are subject to obligations such as publishing a contact point to receive and distribute notifications. Usually, these services create a random and unique email address for their customers, using their brand suffix. This is entered into the Private Registration Address field of the WHOIS record. Thereafter, when messages are sent to that email address, these services forward the messages to the email address customer listed in their internal registration data. In our dataset, these services are observed for both domain and IP WHOIS records. We consider an email addresses to be privacy protected, if the suffix of the address corresponds to one of 17 privacy-protection services we identified.
- x_6: **Free Email**: We consider an email address from a free email provider when the domain name of the email address matched with a list of free email providers (publicly available in [67]). This list contains both currently active and defunct providers. We hypothesize that having a free email account reduces the probability of a bounce, because the same email address could also be being used as a personal email.

We used these variables to model the probability that a notification bounces. A multivariate logistic regression analyses was carried out to assess the influence of each variable. Logistic regression does not restrict the type of variables that can be used. They can be continuous, discrete or a combination of the two. Additionally, the variables do not necessarily have to have a normal distribution. The binary logistic regression equation is:

$$logit(\pi_b) = log\left[\frac{\pi_b}{1-\pi_b}\right], \tag{3.1}$$

where π_b is the occurrence probability of an email to bounce within the range [0, 1] and can be estimated as:

$$\pi_b = \frac{exp(\beta_0 + \sum_i \beta_i x_i)}{1 + exp(\beta_0 + \sum_i \beta_i x_i)}, \tag{3.2}$$

where x_i $(i = 1, \ldots, 6)$ refers to the explanatory variables; β_i is the partial regression coefficient; and β_0 is the intercept. $exp(\beta_i)$ is an odds ratio, which mirrors the strength of the correlation between the explanatory variables and the bounce probability. When $exp(\beta) > 1$, a positive correlation exists between the variables and the occurrence probability. When $exp(\beta) < 1$, a negative correlation exists. When $exp(\beta) = 1$, the variables are not correlated with the event.

The results are presented in Table 3.5. All variables have a significant effect on the bounce rates.

Table 3.5: Coefficients of the logistic regression model for email bounce occurrence

	Dependent variable:
	bounced
x_1 : SOA	0.794***
	(0.061)
x_2 : whoisDom	−1.752***
	(0.091)
x_3 : whoisIP	−2.333***
	(0.175)
x_4 : selfGenRFC	1.929***
	(0.067)
x_5 : whoisprotection	0.698**
	(0.272)
x_6 : freemail	−1.109***
	(0.227)
Observations	5,051
Log Likelihood	−2,175.878
Note:	*p<0.1; **p<0.05; ***p<0.01
	Standard errors in brackets

Coefficients in logistic regression models can be interpreted as odds-ratios. By calculating the odd ratio from the estimated coefficients, we observe that:

- Contacting affected parties using self-generated email addresses based on RFC stan-

dards increases the odds of delivery failure by 588% (odds ratio : 6.88, confidence interval: [6.05, 7.86]).

- Contacting resource owners by using addresses from the SOA record `RNAME` field increases the odds of delivery failure by 121% (odds ratio: 2.21, confidence interval: [1.96 , 2.49]).
- Using the abuse email field of `IP WHOIS` records for notifications decreases the odds of bouncing by 90% (odds ratio: 0.09, confidence interval: [0.06 , 0.13]).
- Using a privacy or proxy services doubles the probability of the email to bounce (odds ratio: 2, confidence interval: [1.15 , 3.36]).
- Contacting addresses from free email providers, as found in `WHOIS` records and SOA `RNAME`, decreases the bounce occurrence by 67% (odds ratio: 0.32, confidence interval: [0.20 , 0.50]).
- Using an addresses gathered from domain `WHOIS` records decrease the bounce probability by 82% (odds ratio: 0.17, confidence interval: [0.14 , 0.20]).

As we hypothesized, contacting affected parties via addresses from `WHOIS` records reduces the odds of a bounce. If this address is from a free email providers, this further reduces the bounce probability. On the other hand, recipients that are behind privacy-protection services have a significantly higher bounce rate, even though these emails are also gathered from `WHOIS` records.

By far the worst performing in terms of deliverability are self-generated email addresses compliant with RFC recommendations. This mainly indicates that very few nameserver operators and network providers actually follow the recommendations. Many domain owners and DNS services providers (or owners) do not correctly format SOA records, nor integrate mailboxes for security and operational needs.

We assess the goodness-of-fit of our model by calculating the Receiver Operating Characteristic Curve (ROC). The ROC summarizes the model performance between true positive (TP) and false positive (FP) error rates. Figure 3.7 shows the ROC curve of the model. The area under the ROC curve (AUC score) adds to a combined sensitivity and specificity of 85%. This indicates a good discrimination power of our model when predicting an email will bounce based on the six explanatory variables.

3.4.2 Modeling Remediation Occurrence

We now turn to remediation. We model the chance of remediation as a function of certain features of the nameserver. We derived five variables that might affect remediation:

- **Communication Channel**: This categorical variables represents the type of channel used to reach the nameserver operator. In our experiment we had three different communication channels:

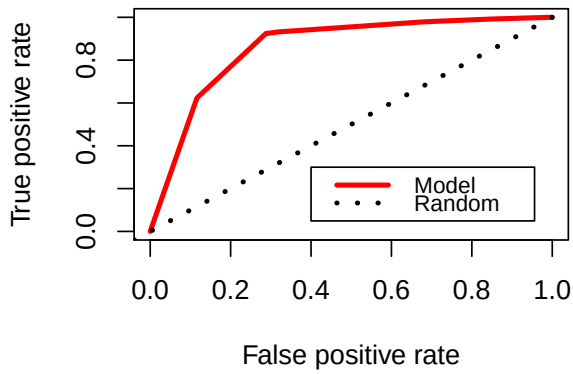


Figure 3.7: Logistic regression diagnostic with ROC curve

  - x_1: **Direct Channel**: This channel was used during the first campaign as the recipient of the notification was the nameserver operator.
  - x_2: **Indirect Channel Through Domain Owner**: This channel was used during the second campaign where the customers of the nameserver operator were the recipients of the notifications.
  - x_3: **Indirect Channel Though Network Operator**: This channel was used during the third campaign where the notification were sent to the handler of the nameserver IP address.
- x_4: **Number of Vulnerable Domains**: Count of vulnerable domains under a specific nameserver as seen in passive DNS data available in DNSDB.
- x_5: **Number of Domains**: Total number of domains under a given nameserver as seen in passive DNS data available in DNSDB.
- x_6: **Domain Popularity**: Logical variable set to `TRUE` when one or more domains under a specific nameserver are in Alexa's one million top-ranked domains.
- x_7: **Link to Demonstration Website**: Logical variable set to `TRUE` when the recipient of the notification received the notification with the link to the demonstration tool.

We used a logistic regression model to estimate the probability of remediation from the aforementioned explanatory variables. Table 3.6 shows the results. We performed a stepwise inclusion of variables per model. As we move from the initial model to the fifth, we aim to improve the accuracy of model's remediation probability prediction. The discrimination power of the model increased as we added new explanatory variables. The fifth model is the final model we use to explain the remediation occurrence.

Table 3.6: Coefficients of the logistic regression model for nameserver remediation occurrence

	Dependent variable:				
	Nameserver Remediation Occurrence				
	(1)	(2)	(3)	(4)	(5)
x_1 Direct Channel	1.591***	1.456***	1.470***	1.463***	1.465***
	(0.291)	(0.300)	(0.300)	(0.300)	(0.300)
x_2 Indirect Channel$_1$	1.012***	0.867***	0.885***	0.866***	0.859***
	(0.294)	(0.304)	(0.305)	(0.305)	(0.305)
x_3 Indirect Channel$_2$	0.959***	0.817***	0.803***	0.803***	0.786**
	(0.295)	(0.305)	(0.306)	(0.306)	(0.307)
x_4 Number of Vulnerable Domains					0.004***
					(0.001)
x_5 Total Number of Domains			0.0001**	0.0001**	−0.0001
			(0.0001)	(0.0001)	(0.0002)
x_6 Hosting Popular Domains				0.632***	0.607***
				(0.194)	(0.195)
x_7 Link to Demonstration Website		0.252*	0.236*	0.225*	0.216
		(0.132)	(0.133)	(0.133)	(0.133)
Constant	−3.676***	−3.676***	−3.689***	−3.738***	−3.731***
	(0.271)	(0.271)	(0.271)	(0.272)	(0.272)
Observations	3,956	3,956	3,956	3,956	3,956
Log Likelihood	−965.880	−964.047	−958.999	−954.294	−950.437

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

In the model, the only non-significant factors are the nameserver size and whether the recipient received the link to the demonstration website. We interpret the coefficients as odds ratios. This provides us with the following observations:

- Having nameservers that include popular domains increase the odds of remediation by 83% (odds ratio: 1.83, confidence interval: [1.23 , 2.65]).
- An increase in the number of vulnerable domains on a nameserver has no effect on it being remediated (odds ratio: 1.00, confidence interval: [1.00 , 1.00]).
- Direct notifications increase the odds of nameserver remediation by 332% (odds ratio: 4.32, confidence interval: [2.47 , 8.10]).
- Notifications to domain owners increase the odds of nameserver remediation by 136% (odds ratio: 2.36, confidence interval: [1.33 , 4.45]).

- Notifications to network operators increase the odds of nameserver remediation by 119% (odds ratio: 2.19, confidence interval: [1.23 , 4.15]).

As we see from the results, the size of the nameserver (number of domains) and sending the notification with a link to the demonstration site did not significantly influence the remediation occurrence in the final model. They were already only weakly correlated in the prior models, which explains the sign flips and changes in significance. Although the number of vulnerable domains on a nameserver was statistically significant, it has very small effect on the odds of remediation.

These results also indicate that direct notifications made the highest impact across all variables. It increased the odds of remediation by 332% compared to 136%–119% for those notifications sent through an indirect channel.

Similarly to the previous model, we assess the goodness-of-fit by calculating the ROC curve and the computing the AUC value. Though the ROC curve for the model (see Figure 3.8) shows that model can predict slightly better than the random model (with 69% AUC score), it has poor predicting capabilities.

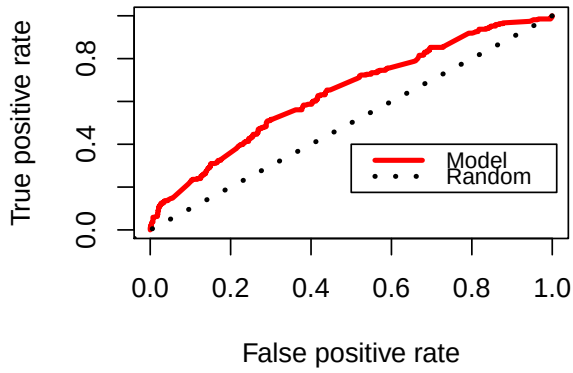


Figure 3.8: Logistic regression diagnostic with ROC curve

3.5 Reactions of recipients

We observed the reactions of recipients through their email replies and the results of an anonymous survey. All of our notifications included a link to the survey. We received 25 survey responses. This renders the survey useless in terms of understanding the population of recipients. We do discuss the results as anecdotal data that helps to think about how such scans and notification campaigns might be perceived by the affected parties. We received

23 responses to our notifications via our contact email. In following section we analyze reactions of recipients.

3.5.1 Survey Responses

The survey was anonymous and each question was optional (see Appendix B.4 for more details). The questions were slightly tailored to the different campaigns and treatment groups. Each survey consisted of 10 questions, with an extra question for the recipients contacted via the indirect channels, asking them whether they could fix the problem by themselves or not.

The survey began by asking demographic-type of questions on the type of organization where the recipient worked and the size of the organization. Next, we asked them whether they had taken any action before getting our notification and whether they were planning to take any action after the notification. These questions are followed by another two questions to learn whether the recipients found it acceptable to scan their nameserver and notify them about the vulnerability. The survey for recipients that received the link to the demonstration website asked about the effectiveness of the demonstration website. Recipients of the conventional notification were asked whether it would have been useful to be provided with a site to safely test and demonstrate the vulnerability. Recipients were also asked whether we notified the right contact point and whether they would like to receive future notifications. At the end of the survey, respondents were given an open question and asked to tell us they wanted about the scans, notifications or any other issue related to this research or to security notifications in general.

We summarize the results in Table 3.7. In the first campaign, we received 11 responses, 5 of them from the demonstration group and 6 of them from the conventional content group. Most responses were from hosting providers. The rest of the responders were representatives of DNS service provider, software and gaming company, content delivery network and government. In this campaign, the majority of the responders were from small and medium size organizations and only two responses were from large organizations. In second campaign, where we contact the domain owners, we received 9 responses. Of these, 5 of them belonged to the demonstration group and 4 to conventional notification group. Similarly to campaign 1, small and medium size of organizations were represented more than large organizations. In the third campaign, we received 5 survey responses, 3 of them from the demonstrative notification group and 2 of them from the conventional notification group. The majority of the responders in 3 campaign had large number of employees.

Surprisingly, 9 responders in first and second campaign indicated that they had previously attempted to remediate the problem. After the notification, nearly all responders were planning to remediate the problem. The majority of the responders found our scans and notifications acceptable and they were open to future notifications. Moreover, 23 responders indicated that we reached the correct contact.

In the first campaign, when we asked about the usefulness of the demonstration web-

site, 3 respondents found it very useful and 2 of them found somewhat useful. No one found it not useful. Similarly, the conventional notification group was asked whether it would had been useful if we had provided a demonstration website, 4 of them replied that it would have been very useful and 2 of them replied somewhat useful. In the second campaign, the demonstration website was found very useful for 3 respondents, one did not go the site and one find it not useful. Moreover, half of the responders who belonged to the conventional notification group indicated that providing link to a demonstration website would have been very useful and the other half indicated that providing a link to a demonstration website would have been somewhat useful. In the third campaign, one of the respondents find the demonstration website very useful, other find it somewhat useful and the rest did not go to the site. The respondents for the conventional notifications indicated that providing a link to a demonstration website would have been very useful.

Since in the second and third campaign an indirect channel was used, we wanted to know whether any of the responders in these groups were maintaining the server by themselves. According to these responses, 8 (out of 14) respondents were capable of maintaining the vulnerable server.

3.5.2 Email Responses

Throughout the study, we had 23 human replies to our emails (see Table 3.8). Two emails stated that the servers in question did not belong to the recipient. Five emails were negative. Two people complained about the scans, one threatened to sue, one claimed to have reported us, not mentioning to whom, and one shared a rather unimaginative insult.

3.6 Related work

The effectiveness and feasibility of security notifications has recently become a major concern [21]. Several researchers have begun investigating how security notifications can expedite vulnerability remediation. Li et al. studied the aspects of vulnerability notifications that could increase the vulnerability remediation rates [43]. Their study focused on who to notify and how much information does needs to be included in the notifications. They found that security notifications addressed directly to the owners of the vulnerable resources promote faster remediation than those sent to national CERTs and US-CERT. In addition, their study revealed that detailed vulnerability notifications increased the vulnerability remediation rate compared to terse vulnerability notifications. Stock et al. investigated the feasibility and efficacy of large-scale notification campaigns [44]. Their findings indicated that vulnerability notifications increased the patching rate compared to those that are not notified. However, overall patching rate was marginal. Prior to these work, Kührer et al. conducted a collaborative notification campaign with the CERT/CC and Cisco to notify the network providers and owners of equipment running vulnerable NTP servers [42].

Table 3.7: Survey responses

Survey Responses	Campaign 1		Campaign 2		Campaign 3	
	Demonstration	Conventional	Demonstration	Conventional	Demonstration	Conventional
Number of participants	5	6	5	4	3	2
Organization type	Hp:5 S/G firm: 1	HP: 2 DNS:1 CDN:1 Government:1	DO: 2 Organization:3	DO: 1 Organization:3	ISP: 2 Private org:1	ISP:1 HP:1
Size (if organization)	1: 1 25-99: 2 100-499: 1 500-999:1	1: 3 2-24: 1 25-99:1 500-999:1	1-24:4 25-99:1	25-99:1 1000+: 2	100-499:1 1,000+:2	2-24:1 100-499:1
Taken Prior Actions	3/2	2/4	1/4	3/1	0/3	1/1
Now Taking Action	4/1	6/0	5/0	4/0	2/1	1/1
Acceptable to Scan	5/0	6/0	3/2	4/0	3/0	1/1
Acceptable to Notify	5/0	5/1	5/0	4/0	2/1	2/0
Demonstration website useful (if provided one)	Very useful:3 Somewhat useful: 2	Very useful:4 Somewhat useful: 2	Very useful:3 Didn't go: 1 Not useful: 1	Somewhat useful: 2 Very useful:2	Very useful:1 Somewhat useful: 1 Didn't go:1	Very useful: 2
Future Notifications	5/0	6/0	4/1	4/0	2/1	2/0
Correct Contact	4/1	6/0	5/0	4/0	2/1	2/0
Can maintain the nameserver	–	–	3/2	3/1	1/2	1/1

Table 3.8: Email Responses

Human Responses	Campaign 1		Campaign 2		Campaign 3	
	Demonstration	Conventional	Demonstration	Conventional	Demonstration	Conventional
Positive Remark	0	1	4	1	5	1
Negative Remark	0	1	3	0	0	1
Neutral Remark	1	0	0	0	1	0
False Positive Notification	2		0		1	

They observed a 92% reduction in vulnerable servers, from 1.6 million to 126,000 in under three months. Regarding the high-profile disclosure of the OpenSSL Heartbleed bug, Durumeric et al. notified operators of detected vulnerable hosts and found that the rate of patching for notified group was 47% higher than the control group [41].

More recently, a number of papers have also started to investigate impact of abuse notifications. Li et al. described in detail the impact of security notifications on 761,935 infected websites that were detected by Google Safe Browsing and Search Quality [33]. They discovered that direct notifications to webmasters via Google Webmaster Console increased the likelihood of cleanup by over 50% and reduced the infection lifetime by 62%. Furthermore, in Chapter 2, we investigated the impact of sender reputation in abuse reports. We found no statistically significant difference between the abuse notifications of senders with varying level of reputation, suggesting that the sender email address does not matter greatly when responding to abuse reports. However, we observed that the notifications resulted in

better cleanup than not notifying.

In another previous study, Vasek and Moore conducted an experimental study on malicious URLs submitted to the StopBadware community feeds [32]. They found that abuse notifications sent with detailed information on the compromise are cleaned up better than those not receiving a notice. Surprisingly, they found no difference between the cleanup rates for websites receiving a minimal notice and those not receiving any notice at all.

In two other studies, researchers experimented with web-based malware in hosting services. In a first study, Nappa et al. sent abuse reports to providers hosting 19 long-lived exploit servers [35]. Only 7 out of 19 providers took action towards cleaning up the malicious servers. In a second study, Canali et al. set up vulnerable webservers on 22 hosting services [34]. They then compromised the webservers and sent out notifications to all hosting providers after 25 days had passed. Approximately 50% took action, generally suspending access to the websites. To ensure that the notifications were actually being read and not simply being acted upon without evidence, false abuse reports were also sent, resulting in 3 of the 22 providers suspending an account without actual evidence. This demonstrates the pitfalls in investigations on abuse reports.

Moreover, Gañán et al. studied how different forms of notifications affected lifetime of ZeuS command and control servers provided by ZeuS Tracker, Cybercrime Tracker and a private company [36]. While ZeuS Tracker and Cybercrime Tracker present a publicly accessible dynamic webpage that displays ZeuS malware command and control servers, the private company did not publicize any of the detected command and control servers. Research concluded that publicized command and control servers were mitigated 2.8 times faster than the ones that were not publicized.

Furthermore, in another study Vasek et al. studied impact of the incident data sharing among Internet operators [38]. Their study concluded that sharing abuse data has a swift effect of cleaning the reported malicious URLs.

Finally, with respect to spam, a quasi-experiment by Tang et al. used two blocklists to compile a large source of e-mail spam and publish aggregated measures on SpamRankings.net[37]. They then published the results for a treatment group and withheld results for a control group, observing a 15.9% reduction in spam among the treated group. Rather than notifying individual hosts in order to remediate infections, the researchers' strategy relied on public shaming. The study indicates that reputation effects could provide an incentive for intermediaries to cooperate in remediating abuse on their networks.

3.7 Conclusions

We succinctly state the main results and discuss what they tell us about improving the effectiveness of vulnerability notifications.

3.7.1 Reaching affected parties at scale

In light of the rise of large-scale vulnerability scanning, our most sobering result is that there is no good mechanism of getting this wealth of information to the relevant entities. Most of our notifications bounced. Contact information is extremely unreliable. RFC standards, which might help make the system more robust, are widely ignored. There is a large and growing discrepancy between our ability as a community to collect information and our ability to make this information useful for those under threat.

It is not clear where to go from here. One could find a bit of solace in the fact that network operators did much better in terms of being reachable. Should we direct our notifications more to them? This will surely overload them. Their IP address space may be filled with hundreds, thousands, or even tens of thousands of affected systems. Another disadvantage is that in terms of remediation, this path was not more effective. Perhaps they are too far removed from the resource owner or operator to really do anything, except forward the notification. This is already a non-trivial task, which requires dynamically mapping notifications to the relevant customer in their network.

What else could be done? One option is to move away from email as the main notification medium. There are other options that are likely to be more effective, such as automated feeds, APIs, or sharing data within specific communities. For instance, Kührer et al. issued notifications (about systems vulnerable to abuse in NTP DDoS amplification attacks) to key organizations such as abuse team contacts at CERTs, security data clearinghouses [42].This indirect approach proved very effective: 92% of the amplifiers were remediated in three months.

The problem with these alternative information sharing mechanisms is that they are typically based on opt-in. Given that many of the affected parties in our studies didn't even set up a correct SOA record or put a working email address in their WHOIS record, it is difficult to be optimistic about any information sharing mechanism that requires an active effort on the side of the recipient. This question will have to be picked up by the industry, CERT and CSIRT community, Regional Internet Registries and others. Getting perfect reachability is unlikely to happen any time soon, but it should definitely be possible to improve beyond the current sorry state of affairs.

3.7.2 Incentives for remediation

While notifications did lead to more remediation than in the control groups, the overall remediation rates were low. Now, one issue is that not all vulnerabilities need be remediated. This fact is under-appreciated by the well-meaning efforts to increase vulnerability scanning and notification. Remediation represents an economic tradeoff and the outcome depends on the threat model of the affected party. This issue is undoubtedly also in play among the recipients of our notification. That being said, to offer total control over your DNS records to anyone on the Internet seems like an obvious problem that should be fixed. Some potential fixes, or perhaps it is better to call them workarounds, can be applied in a

relatively simple manner. So why aren't they? Is it a lack of awareness? Incompetence? Lack of resources? The truth is: we don't know.

Security economics has taught us that systems are particularly prone to failure when the actor protecting it does not suffer the full cost of failure. Perhaps the incentive of the nameserver operator is too weak, as the abuse would impact the domain owner first and foremost. For this reason, we investigated if the incentive structure for remediation was stronger when we contacted the domain owners, who could then request the remediation from their provider, leveraging the commercial incentive of the latter party.

Our study found that this mechanism does not lead to better remediation. If remediation is a matter of incentives, then this indirect path either has equally weak incentives, or the stronger incentives are neutralized by the higher friction in the process towards remediation. In any case, the conclusion is that we need to look for other ways to improve the incentives. Some have pointed to reputation effects – a.k.a. naming, praising and shaming – as potentially effective [68].

3.7.3 Usefulness of the Demonstration Website

Another part of the incentive puzzle is more behavioral in nature. Recipients often need to triage notifications, and this will only increase in the age of large-scale vulnerability scanning. In this process, being able to asses the credibility, trustworthiness and criticality of the issue, might nudge recipients towards action. We tested whether mitigation improved when a website was provided with a live demonstration of the vulnerability for the recipient's domain or nameserver. The short answer is: no, remediation did not improve. The handful of responses to our survey do suggest, however, that the demonstration was helpful. So the bottleneck appears to be to get recipients to actually visit the site via a notification message. This is a complicated issue, as it triggers all kinds of overtones of phishing and other red flags for security-conscious persons. One way forward might be to host such a site at a trusted node in the network, such as the national CERT. Future work will have to test whether this has a more observable impact.

CHAPTER 4

Measuring effectiveness and usability of quarantining compromised users in walled gardens

In the fight to clean up malware-infected machines, notifications from Internet Service Providers (ISPs) to their customers play a crucial role. Since stand-alone notifications are routinely ignored, some ISPs have invested in a potentially more effective mechanism: quarantining customers in so-called walled gardens. In this chapter, we present the first empirical study on user behavior and remediation effectiveness of quarantining infected machines in broadband networks. First, we explain how walled garden notifications issued by the ISP and types of release mechanisms that can be used by the quarantined users. Then, we investigated the impact on infection type and release mechanisms used by the quarantine users on malware remediation speed. Lastly, we studied the reactions of quarantined users to get a better sense of the experience of the end users.

4.1 Introduction

Fighting the scourge of malware-infected end user machines is an ongoing challenge that involves many different actors, from software vendors, incident response organizations, antivirus vendors, network operators and, last but not least, the end users themselves. Some efforts are more focused on preventing infections, others on remediation – i.e., cleaning up the compromised hosts. In the context of cleanup, the role of Internet Service Providers has become more salient over time, as it became clear that many end users struggle to detect and remediate infections. The ISPs are a critical control point providing the infected machines with access to the rest of the Internet. In the past 5-10 years, a range of best practices and code of conducts have been published by leading industry associations [69, 70], public-private initiatives [71, 72] and governmental entities [73, 74]. These documents share a common set of recommendations for ISPs around educating customers, detecting infections, notifying customers, and remediating infections.

The effectiveness of these best practices is disputed. When the U.S. National Institute of Standards and Technology (NIST) was developing its own guidance on 'Models To Advance Voluntary Corporate Notification to Consumers Regarding the Illicit Use of Computer Equipment by Botnets and Related Malware', it considered using the Australian iCode as an example [75]. The SANS Institute and other stakeholders criticized this idea, arguing the Australian code had not managed to significantly improve cleanup rates of infected users [75]. Academic research has also questioned the effectiveness of these efforts [76, 77].

There are a variety of reasons for the limited impact of botnet remediation efforts by ISPs. At the core, however, is a usability problem: notifying customers that one of their machines is infected does not translate into actual cleanup. As we know from other areas in security, notifications are routinely ignored, especially if the step towards action is complicated and disrupts ongoing activities.

The lack of effectiveness of mere notifications has led some of the more security-minded ISPs to adopt what is arguably the most costly measure: putting infected customer machines into a quarantine network, also known as a 'walled garden', which only gives access to a small set of white-listed sites. Users are required to perform cleanup to get their connection restored – i.e., to be released from the walled garden. While the use of walled gardens is identified as a security best practice [78], it is also controversial. The ITU's Anti-Botnet Toolkit cites 'technical, financial, legal and customer satisfaction-related disincentives' that may be raised by an ISP [79].

Quarantining infected users is contested, but also one of the few measures that could improve cleanup rates and help end users to remediate and secure their machines. Remarkably, there has been no publicly available study on the effectiveness of walled gardens. Do they actually help end users to clean up? How often do users get reinfected? How much time do users spend in quarantine? How much support do they need? How much pushback do ISPs face from their users?

We present the first empirical study on the usability and effectiveness of walled gardens as a notification and remediation mechanism. We analyzed 6 months of data (April-October 2017) from a real-world implementation of a walled garden at a medium-sized ISP that we collaborated with. The ISP is a market leader in its home market that serves retail broadband to several million customers. The ISP took $1,736$ quarantining actions involving $1,208$ retail customers. In collaboration with the ISP, we correlated these quarantining actions with independent observations from botnet sinkhole data to track remediation success. We also analyzed anonymized communications with quarantined users. In combination, these datasets allow us to estimate cleanup rates, recidivism rates, and user engagement with the walled garden environment.

In short, we make the following contributions:

- We present the first empirical study of a real-world 'walled garden' system to notify and quarantine end users with malware-infected machines – a widely-recognized security best practice for ISPs.
- We measure the effectiveness of the walled garden notifications in terms of end user

cleanup efforts and find that the majority of users spend a relatively short time in quarantine, while still successfully removing the infection.

- We provide insight into the experiences of users by analyzing their communication with ISP employees and find that a fraction of them are frustrated about their access being cut off. This is especially true for users who turn out to operate business services over their consumer broadband connection.

The rest of this paper is structured as follows. Section 4.2 reviews prior work. Section 4.3 outlines the properties of walled garden systems and Section 4.4 presents the data collection methodology. Next, Section 4.5, we shed light on the effectiveness of the real-world walled garden and relationship between cleanup success and other factors. Section 4.6 presents key insights gathered from communications. Section 4.7 presents the ethical considerations and Section 4.8 discusses the limitations of the study. We conclude by covering the main lessons learned for the use of walled garden systems in securing end-user machines.

4.2 Related Work

As far as we are aware, there is no prior work on the effectiveness of notifying end users in an access network and asking them to clean up malware infections on their machines. Here, we briefly survey four related areas of work. The work on abuse and vulnerability notifications has studied similar mechanisms , but typically with a different type of end user, namely webmasters, server admins and network operators, not home users. This makes the effectiveness of those mechanisms difficult to compare with malware notifications and cleanup by consumers. Another area of related work concerns the design of the notifications and warnings for regular end users. These notifications and warnings are mostly meant to prevent compromise, trying to steer the user back to safety. In contrast, we study a notification mechanism where the action is not avoiding danger, but dealing with the damage that has already occurred. Also, the action required of the user in case of compromise is not a single decision for or against a potentially dangerous action, but the execution of a rather complicated set of steps to resolve the incident that has already manifested itself. Finally, there is related work that studies whether and how end users understand the security situations they face and how they behave in those contexts. In our study, we do not observe the users directly, nor elicit their thoughts about the situation, but we do have data on some of their actions, as well as some visibility into their experiences through their communications with the ISP.

4.2.1 Abuse notifications

A range of studies has focused on if and how abuse notifications can expedite cleanup of compromised websites. Notifications can be sent to the affected owners of the site or to

their hosting provider. An early study by Vasek *et al.* [32] indicated that more verbose abuse notifications to hosting providers resulted in higher cleanup rates than notifications with minimal information. In Chapter 2, we found that around half of all compromised sites got cleaned up after a notification to the hosting provider. The reputation of the sender of the notifications had no observable impact on the cleanup rate. Li *et al.* [33] showed that direct notifications to webmasters via Google's Webmaster Console increased the likelihood of cleanup by over 50%. They report that 6.6% of sites cleaned up within a day of detection, 27.9% within two weeks, and 41.2% within one month. In a qualitative study, Canali *et al.* [34] set up vulnerable web servers on 22 hosting services, ran different attacks on them that simulated infections and then notified the providers about these attacks. Only one hosting provider notified their customers about a potential compromise of their website after the first notification and only half of the providers after the second notification. Additionally, around 13% of the notified providers warned the user of being compromised upon receiving abuse notifications.

4.2.2 Vulnerability notifications

Various studies have looked into the feasibility and efficacy of vulnerability notification mechanisms. For example, Kührer *et al.* [42] issued notifications to administrators of vulnerable Network Time Protocol (NTP) servers, in collaboration with CERTs, clearinghouses and afflicted vendors. Though their study lacks a control group to assess the impact of the campaign itself, they found that 92% of NTP server were remediated in 13 weeks. Stock *et al.* [44] studied large-scale vulnerability notification campaigns and found that only around 6% of the affected parties could be reached. Of that small fraction, around 40% were remediated upon notification. Similarly, in Chapter 3, we concluded that the deliverability of email-based notifications was very poor. They proposed searching for other mechanisms. Stock *et al.* [45] later tested the effectiveness of other channels such as postal mail, social media, and phone and concluded that the slightly higher remediation rates of these channels do not justify the additional work and costs.

4.2.3 Design of notifications and warnings

A large body of literature explored user responses to different types of security notifications and warnings, focusing on why users ignore warnings and how this could be avoided. A study conducted by Krol *et al.* [80] showed that users' misunderstanding of warnings and notifications is a reason for ignoring them. Almuhimedi *et al.* [81] studied user reactions to Google Chrome malware warnings. Up to half of the warnings were ignored under certain circumstances. Some users confused the malware warnings with SSL warnings. Sunshine *et al.* [82] examined users' reactions to existing and newly designed SSL warnings and suggested that, although existing SSL warnings can be improved, minimizing the use of SSL warnings by blocking users from making insecure connections proves to be more effective.

Finally, Mathur *et al.* concluded that one of the reasons why users ignore software updates is that updates regularly interrupt users who often lack sufficient basic information to decide whether or not to update [83]. A closely related topic is the problem of habituation of users to ignore warnings after they have learned that this does not seem to cause any harm [84, 85]. Bravo-Lillo *et al.* tested the effectiveness of user-interface modifications to draw users' attention to the most important information required for decisions [86, 87].

4.2.4 End user security behavior

Multiple studies have demonstrated that end users have difficulty securing their computers, either because of lack of knowledge or ignoring security advice that is hard to understand. In a study conducted by Wash *et al.* [88] on how users perceive automated software updates, the authors observed that the majority of users do not correctly understand the automatic update settings on their computer and cannot manage software updates the way they intend to. This mismatch between intention and behavior frequently led to computers being more or less secure than intended. Fagan *et al.* [89] studied user motivations regarding their decisions on following common security advice (i.e., update software, use password manager, change passwords) and concluded that the majority of users follow the usability/security trade-off. Finally, Forget *et al.* [90] developed a Security Behavior Observatory to collect data on users' behavior and their machine configurations. Their findings highlighted the importance of content, presentation, and functionality of security notifications provided to users who have different expertise, expectations, and computer security engagement.

4.3 Walled Garden

The concept of a "walled garden" stems from the early days of the web, when ISPs implemented closed networks to control the applications, content and media that their subscribers could access. Some ISPs extended the capabilities of these networks to exclude rival content from the heavily curated garden. This model has all but disappeared.

These days, walled gardens are a method to notify subscribers about malware infections and restrict their access to the Internet while infected, so as to protect the infected user from further harm as well as preventing the user's machine from harming other users or networks. More precisely, a walled garden is a quarantined environment that restricts the information flow and services of an end user inside a network. Besides keeping the infected users safely in quarantine, the walled garden also plays an important role in informing the user. While the user tries to browse the Web, she or he will be redirected to a landing website with information about the type of infection and how to clean it up. Whereas emails or letters with the same content can be ignored relatively easily, this mechanism cannot.

There are different ways of implementing and deploying walled gardens to fight malware infections. RFC6561 [69] describes 2 different types: *strict*, a walled garden environment that restricts almost all services, except those to a whitelist of malware mitigation services; and *leaky*, an implementation that permits access to all Internet resources, except those that are deemed malicious, and ensures access to those that can be used to notify users of infections. In this paper, we focus on a strict implementation, which is what was installed at our partner ISP. A strict implementation is potentially more effective, but also more contested.

The quarantine period of an infected user mainly depends on three different processes: (i) the malware detection process; (ii) the infection notification and quarantining process; and (iii) the release process. The flow chart in Figure 4.1 shows the overall quarantine process in place at our partner ISP. It starts with the ISP realizing that a subscriber is infected and ends with the subscriber leaving the walled garden. The starting point, i.e., the infection detection, is independent of the walled garden environment. Typically, this detection is not based on their own network monitoring, but on third-party notifications, e.g., from botnet sinkhole operators and security intelligence providers. The processing of abuse feeds varies per ISP, ranging from manually checking incoming notifications to highly automated systems that consume the feed and push the relevant incidents into abuse ticketing systems. When certain abuse data fits a predefined policy, on data trustworthiness, timeliness, the affected customer type and other criteria, the ISP places the connection of that particular customer into the walled garden.

In order to leave the walled garden, the customer is requested to provide proof of the cleanup actions that were taken to mitigate the infection. This proof might consist of the log of an anti-virus scan or some description of the steps taken by the user. To facilitate the cleanup, the walled garden can provide access to a range of white-listed services. Typically these services include free antivirus tools and trusted software suppliers. Other white-list entries may be added to protect critical services for the user, such as webmail services and online banking. Thus customers can perform basic remediation steps and communicate with the abuse desk, even though they are quarantined.

After leaving the walled garden, there is no guarantee that the malware infection was actually remediated. There are several reason by which a user could get out of the quarantine network while being still infected. First of all, certain walled garden implementations allow users to self-release at any time. Normally, this option is only available for the first and perhaps second infection event during a specific period of time. When a user is placed in quarantine for a third time, because of a reinfection or because the earlier infection was not actually removed, the option of self-release is no longer available. The quarantine removal can now only be executed by the ISP's abuse or support staff. Second, a user can provide erroneous cleanup proofs. For instance, with an increasing number of connected devices in subscriber networks, it is possible for a non-savvy user to perform cleanup actions on a non-infected device and provide the wrong cleanup proofs to the ISP. It is also possible that advanced malware could remain undetected by common antivirus or removal tools. This

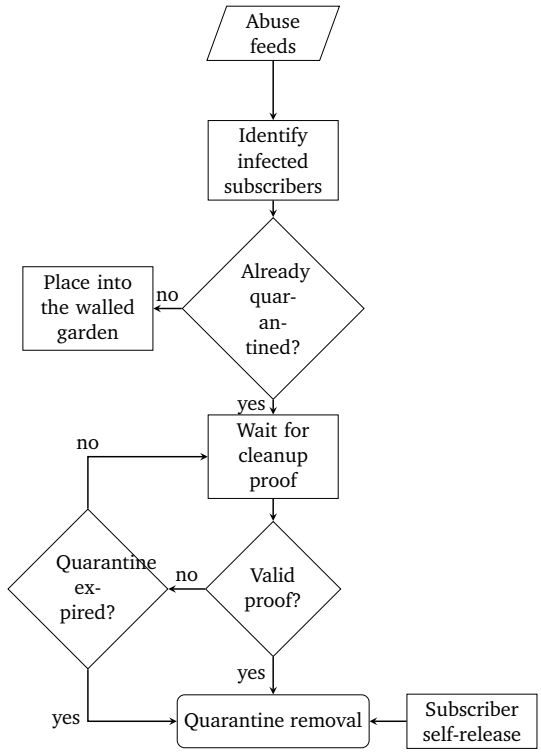


Figure 4.1: Quarantine flow chart

will allow infected users to leave temporarily the walled garden until the same infection is detected again. Third, some walled garden implementations have an expiration period after which any user in quarantine is released. Fourth, and last, ISP staff might decide to release the user without cleanup. Infected users might request to leave the walled garden for other reasons, like an urgent need for certain online services or because the malware infection cannot be remediated while being in the walled garden. The ISP might allow the user to access the Internet to gather a non-whitelisted cleanup tool.

Our study has been conducted on a walled garden environment deployed for the home users of a medium-sized ISP. Their enterprise and mobile customers are not quarantined. The walled garden follows a strict implementation that redirects users to a landing page (see Appendix C.1) and limits the access to a set of 41 white-listed websites, including cleanup tools, antivirus solutions, Microsoft updates, webmail providers and online bank-

ing. Their implementation of the walled garden provides users with two chances to self-release within a period of 30 days. With the third quarantine action, the option to self-release is revoked and the intervention of the ISP's abuse staff is required. After a period of 30 consecutive days in quarantine, the walled garden automatically releases those quarantined customers who did not self-release or contact abuse staff.

4.4 Data Collection

In this section we describe the data that was provided by an ISP to analyze the effectiveness of a particular implementation of a strict walled garden. Our study consists of $1,736$ quarantine events associated with $1,208$ unique subscribers of a medium-sized European ISP's network during a 6 months period. The data was gathered from four different sources that support the ISP's abuse management process: (*i*) abuse feeds providing security incident data to ISPs; (*ii*) walled garden logs recording details of quarantine events in the ISP's network; (*iii*) help desk logs containing the ISP's help desk communication with customers; and (*iv*) abuse desk communication logs providing email exchange between abuse desk employees and customers.

4.4.1 Abuse feeds

In order to detect botnet-related infections, the ISP under study leverages abuse feeds provided by the Shadowserver Foundation. For our analysis, we gathered the Shadowserver botnet reports, collected over a time frame of 9 months between April 10th, 2017 and December 30th, 2017. Three different types of reports are analyzed:

- *Drone Reports*: Drone reports contain detailed information on infected machines discovered through monitoring sinkhole traffic, malicious scans and spam relays. We observed a total of $1,620$ number of malware infected customers in the network managed by the ISP under review.
- *Sinkhole Reports:* Sinkhole reports contain information about sinkhole servers that did not use the conventional bot signatures such as HTTP referrers. Due to lack of conventional bot signatures, many IP addresses mentioned in this reports do not have a specific infection name. During our study period, we observed $1,598$ unique infected users who had a subscription with the ISP under review.
- *Shadowserver's Microsoft Sinkhole:* Microsoft shares via Shadowserver the intelligence gathered from some of their sinkhole servers. Throughout our data collection period, a small number of malicious IP address related to our ISP were captured by Microsoft sinkholes. We only found 8 IP addresses during our study period.

As shown in Table 4.1, we observe a total of $1,620$ unique infected users in the Drone feed, $1,598$ unique infected users in Sinkhole and 8 unique infected users in MS sinkhole

Table 4.1: Infections per feed and quarantined users

	Sinkhole	MS sinkhole	Drone
# infected users	1,598	8	1,620
% quarantined	22%	63%	59%

feeds. Not all of these infections trigger a quarantine action, as Table 4.1 illustrates. There are several reasons why infected users are not quarantined: (i) the user is a mobile or enterprise customer; (ii) the abuse staff decides that quarantining would make matters worse (as in the case of ransomware, where users are by definition already aware of the infection and the lack of Internet access means they might have no viable way to recover their files); (iii) the walled garden environment was undergoing maintenance; and (iv) there are no quarantining actions during the weekend. Figure 4.2 shows the daily number of unique IP addresses seen in the feeds.

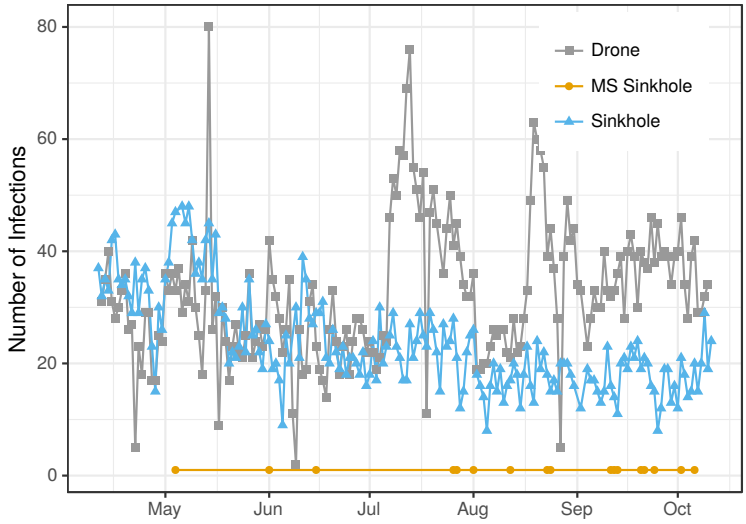


Figure 4.2: Daily unique infected customers per abuse feed

4.4.2 Walled garden logs

During our study period, $1,208$ retail customers were placed into the walled garden based on the abuse feeds provided by Shadowserver. As some customers were quarantined more than once, this corresponds to $1,736$ quarantining events. For each one of these events, several factors were recorded: (i) quarantine time-stamp; (ii) quarantine release mechanism; (iii) quarantine removal time-stamp; (iv) infection type; (v) quarantine event number; and (v) self-release option.

Beside the logs created by the walled garden itself, the quarantined users also have the possibility to submit a form through the walled garden landing page (see Appendix C.2). This form allows users to explain what cleanup actions they have taken, as well as any other feedback they might have. During the study period, $1,575$ forms were received from 831 different infected customers (see Table 4.2).

Table 4.2: Messages and users per communication channel

	Walled garden form	Abuse desk emails	Help desk phone calls
# Users	831	600	468
# Messages	$1,575$	$2,027$	966

4.4.3 Help and abuse desks logs

In addition to the walled garden forms (i), customers can also contact the ISP in other ways. We also collected data on (ii) emails between infected customers and the abuse desk; and (iii) phone calls, store visits and social media chat calls between the help desk and the infected customers. Quarantined customers contacted the abuse desk twice as often as the help desk. Table 4.2 shows that the abuse desk received $2,027$ emails, from 600 unique users while help desk employees reported 966 conversations associated with 468 quarantined users.

4.5 Walled garden effectiveness

We evaluate the impact of the walled garden notification on remediation by looking at the percentage of users that managed to clean the infected machine and at the time an end user remains in the walled garden. We also analyze the relationship between cleanup success and other factors, most notably the type of malware infection, the release mechanism used to get out of the quarantine, and the time spent in the walled garden.

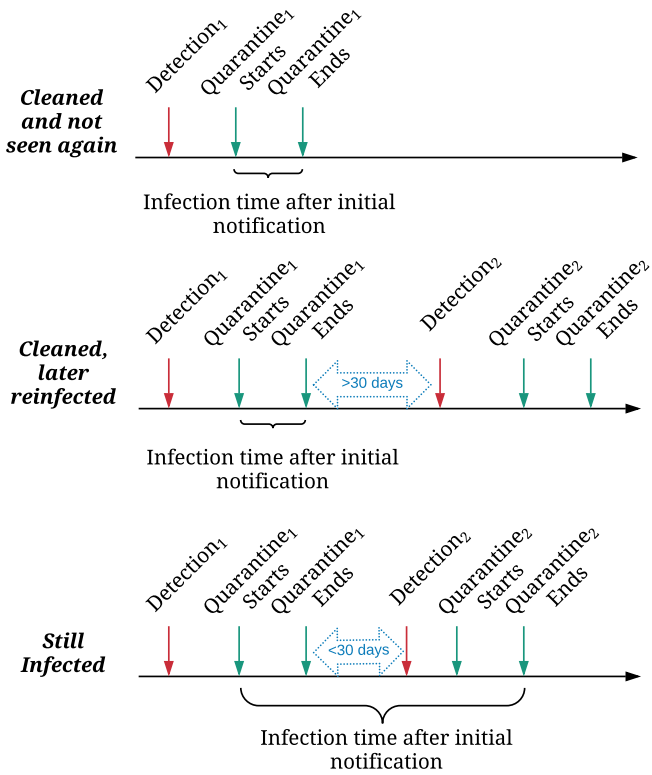


Figure 4.3: Definition of quarantine outcomes

To evaluate cleanup, we distinguish three outcomes when users are released from the walled garden: (i) the user successfully performed cleanup and then stays clean for the rest of the study period; (ii) the user successfully performed cleanup, but the machine is reinfected at a later time in the study period, at least 30 days after the quarantine event; and (iii) the user did not successfully clean up the machine, as evidenced by seeing the offending IP address reported again for the same infection within 30 days of leaving the walled garden.

There is no clear basis for drawing the boundary between a persistent infections and a clean and reinfected machine. Even persistently-infected machines are not seen in the Shadowserver feed every day or even every few days. This depends on a variety of factors,

like the malware type and whether the user even turns on the machine. He or she might be on vacation, for example. We decided to count conservatively in terms of cleanup success and use a long period (30 days) before considering the machine clean. Figure 4.3 shows how these metrics are calculated based on the abuse feeds and the walled garden logs.

There is no clear evidence on where to establish the cut-off point to distinguish persistently infected from clean and reinfected. Figure 4.4 shows the time between consecutive quarantine events. The median time between quarantine events is 4 days. Roughly 70% of the customers who are seen again after being released from quarantine, are seen within 10 days. As gaps in observations are normal for infected machines, this short interval suggests that these machine were probably not cleaned up. After 20 days, the distribution becomes more or less flat with a slow decay. Choosing a cut-off beyond this point only a modest impact on the results. Reinfection rates would change from 16% (day 20 cut-off) to 13% (day 30) to 7% (day 40). As can be seen in the cumulative distribution, around 13% of the users had a gap between quarantine events of 30 days or more – in other words, these are the users we count as cleaned, but later reinfected.

Table 4.3: Cleanup success over number of times in quarantine

Status	Number of times in quarantine						
	#1	#2	#3	#4	#5	#6	#7
Clean and not seen again	830 (69 %)	148 (49 %)	73 (52 %)	18 (35 %)	17 (65 %)	3 (50 %)	2 (67 %)
Clean and later reinfected	51 (4 %)	13 (4 %)	5 (4 %)	2 (4 %)	1 (4 %)	0	0
Still infected	327 (27 %)	142 (47 %)	61 (44 %)	31 (61 %)	8 (31 %)	3 (50 %)	1 (33 %)

4.5.1 Overall remediation rates

In order to understand the effectiveness of the walled garden notifications, we first observe the cleanup and infection rates of the quarantined users after the notifications. We find that 69% of the end users cleaned the infection during their first quarantine event, as shown in Table 4.3. Another 4% of the clean end users got reinfected with the same malware strain at a later point, more than 30 days after the quarantine event. This suggests they did not correctly address the root cause of the infection. The remaining 27% of users were not able to clean the infection.

Most, but not all, users who remained infected or suffered a reinfection, end up in a second quarantine event. Around 20% of them were not quarantined again for a variety of reasons, such as being allowed to leave the quarantine environment to download anti-virus solutions. While this makes the infection show up again in the Shadowserver reports, the abuse desk employees withhold the second quarantining action to see if the user is able to resolve it or not.

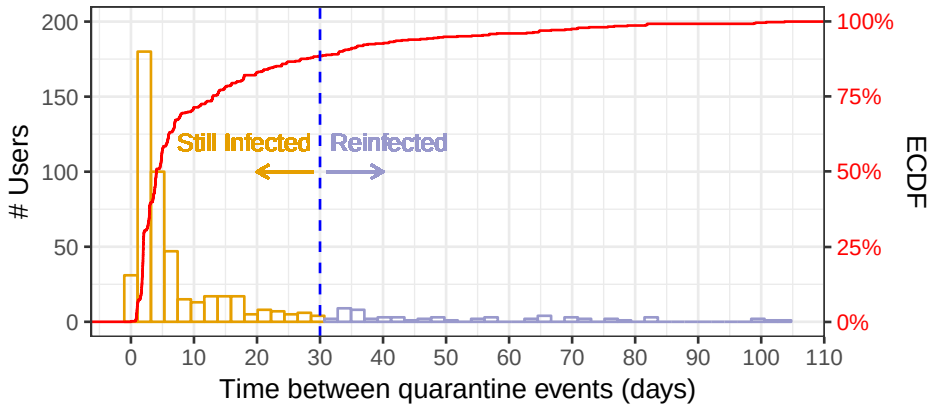


Figure 4.4: Time between consecutive quarantine events

Of those users who ended up in quarantine for the second time, 49% of them now successfully cleaned up the infection. Again, another 4% also cleaned up, but got reinfected later. Around 47% remained infected. We observed that 139 infected end users ended up in quarantine a third time. This time 56% of them managed to remove the infection, including those who got reinfected later on.

In the tail is a group of users, around 4% of all users who ended up in the walled garden during our study period, who suffered four or more quarantine events. At the extreme end, we found three end users who were put into the walled garden seven times over the course of six months.

Next, we explored the infection time after the initial notification for all quarantined end users. Figure 4.5 shows the Kaplan-Meier survival curve of the users' infection and the number of remaining infected users every other day. We find that more than 40% of the infected end users cleaned the infection within a day after initial walled garden notification, 70% within 5 days and only 22% remained after a week. After a month time, only 7% of the users remained infected.

4.5.2 Malware type

We saw that most of the users in quarantine manage to clean up the infection. Does the complexity of an infection influences their success rate and time it takes them to perform the cleanup? Some malware infections might be harder to resolve than others and the white-listed cleanup tools might not always succeed. To understand the influence of the infection type on the cleanup rates, we use the infection names mentioned in the quarantine event

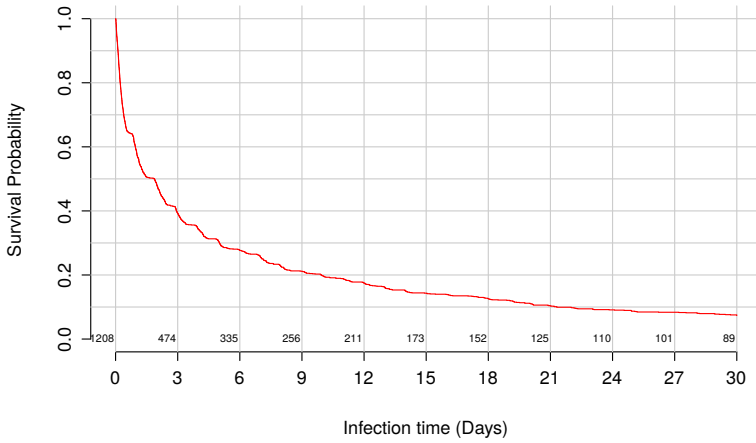


Figure 4.5: Survival curve of the users' infections

logs. The events were triggered by 38 unique infection types. Table 4.4 shows the number of users and quarantined events for the top 10 most frequent infection types, which cover 89% of all the users in our dataset.

Table 4.4: Number of users and quarantine events per malware

Infection	# Users	# Quarantine events
Ramnit	444	675
Mirai	275	410
Nymaim	145	159
Downadup	44	65
ZeroAccess	38	51
Rovnix	34	53
Sality-p2p	34	63
Gozi	21	30
Fobber	20	31
Zeus	20	22

Figure 4.6 plots the survival curves for these infection types during a 30 days period.

We can see significant differences in terms of infection duration for the different infection types (Gehan-Wilcoxon test, $\chi^2 = 58.6$ with $p - value = 2.5e - 09$). For instance, end users infected with "Gozi" managed to cleanup all their infections during a 30 days period. On the contrary, cleanup of the more recent "Fobber" and "Rovnix" malware families was slower than the others. One possible explanation is that the more recent malware is more resistant to the standard cleanup tools linked to in the ISP notification [91].

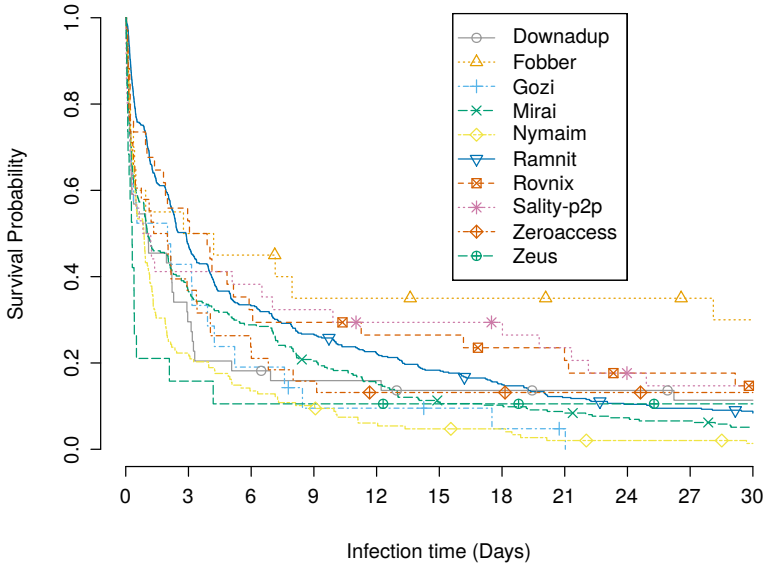


Figure 4.6: Survival probabilities top 10 infection types during 30 days period

4.5.3 Release mechanisms

As we mentioned in Section 4.3, the walled garden contains three mechanisms to release users from the quarantine environment: self-release, assisted release performed by the abuse staff, and quarantine expiry release. Self-release can be used only twice in one month. If this option is disabled, end users can contact help desk employees or abuse desk employees to get out of the quarantine or to ask for more help. However, before releasing the connection back to normal, employees might require evidence of the cleanup action, such as log files of the antivirus software that was used to remove the infection that triggered the notification.

Is there a relationship between the release mechanism and cleanup success? Since

Table 4.5: Quarantine outcomes per release mechanism

Quarantine Event Number	Status	Total # Users	Cleaned, not seen again	Cleaned, later reinfected	Still Infected
1	Self release	805 (67%)	539 (67%)	36 (4%)	230 (29%)
	Assisted	361 (30%)	259 (72%)	11 (3%)	91 (25%)
	Expired	42 (3%)	32 (76%)	4 (10%)	6 (2%)
	Total	1208 (100%)	830 (69%)	51 (4%)	327 (27%)
2	Self release	195 (64%)	84 (43%)	9 (5%)	102 (52%)
	Assisted	102 (34%)	61 (60%)	3 (3%)	38 (37%)
	Expired	6 (1%)	3 (50%)	1 (17%)	2 (33%)
	Total	303 (25%)	148 (49%)	13 (4%)	142 (47%)
3	Self release	17 (12%)	5 (29%)	2 (12%)	10 (59%)
	Assisted	114 (82%)	62 (54%)	2 (2%)	50 (44%)
	Expired	8 (6%)	6 (75%)	1 (13%)	1 (13%)
	Total	139 (12%)	73 (53%)	5 (4%)	61 (44%)

self-release is the fastest and easiest option, one might expect poorer cleanup rates. In the worst case, users simply release themselves without doing anything. To analyze the influence of the release mechanism, we compared the cleanup rates across the first three quarantine actions for all users. As shown in Table 4.5, the first quarantine action ended with 805 users self-releasing, 361 users following assisted release by abuse staff and 42 users were released when the quarantine period expired after 30 days. Of the 805 self-releasing end users, 67% managed to clean the infection. Another 4% also got cleaned, but was later reinfected. In other words, around 71% of all users managed to perform cleanup. Compare this to the cleanup rate of the users who were released by abuse staff after providing evidence of successful cleanup: 75%. These cleanup rates are very close together. Remarkably, self-release does not invite lax security behavior.

Another surprising finding relates to the 3% of users who remained in quarantine until it expired. They had an even higher success rate: around 86%. We do not have an explanation for this. Perhaps these users were fine with only using the white-listed webmail services and, while remaining in quarantine, automated cleanup tools – e.g., Microsoft's Malicious Software Removal Tool, which is downloaded as part of Windows updates – kicked in at some point.

Users who experienced a second quarantine event chose the self-release option in almost the same proportion (64% versus 67% in the first quarantine event). That being said, cleanup rates are not as high as during the first quarantine. In the self-release group, 48% cleaned up successfully (though 5% later got reinfected). In the provider-assisted release,

the cleanup rate is 63%.

During the third walled garden notification period, 82% of the remaining end users ask ISP employees to get them out of the quarantine environment. At this stage, most users no longer get the self-release option, because they were quarantined twice already in one month. Of the users going through assisted release, 54% managed to clean up.

The drop in cleanup rates over successive quarantine events is not large, but might still suggest that perhaps users become habituated and try to get out faster, potentially spending less effort on cleaning and more on getting released. An alternative, and arguable more likely, explanation is that this is caused by selection bias. The users who end up in a second and third quarantine event are likely to be more at risk and perhaps less technically competent. This fits with the fact that with successive quarantine events, the cleanup effectiveness of the assisted-release users become slightly higher compared to the self-release group.

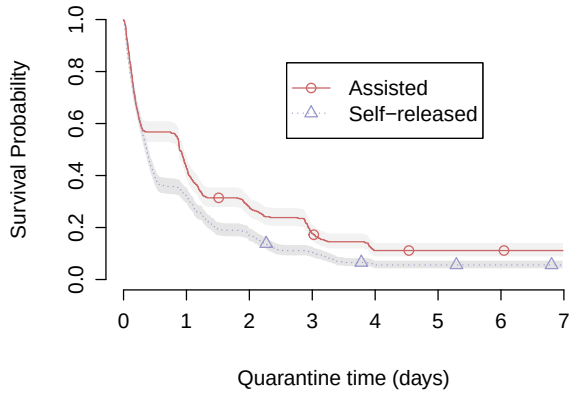


Figure 4.7: Survival probabilities per release mechanism

Figure 4.7 shows the duration of all infections per release mechanism in the form of Kaplan-Meier survival curves. As expected, users that needed assistance to cleanup their infections left the walled garden at a slower rate than the users that self-released. Looking at the speed at which they got removed from the quarantine, we can observe significant differences between these two groups (Gehan-Wilcoxon test, $\chi^2 = 23.1$ with $p - value = 1.5e - 06$). For instance, within the first 2 days in quarantine, 84% of the users that self-released left the walled garden while only 71% of users that needed assistance did so.

4.5.4 Time spent in the walled garden

We now take a closer look at the time users spend in quarantine. Figure 4.8 displays the distribution of the duration of the quarantine events. The majority of quarantine events

lasted less than one day and only 25% of them lasted more than 3 days. A small fraction (57 events) last until they automatically expire.

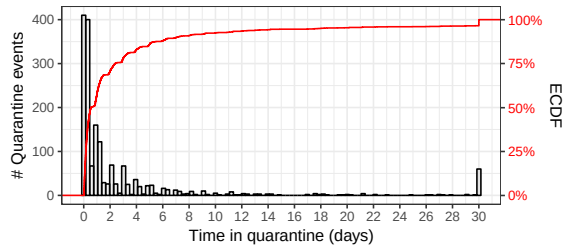


Figure 4.8: Histogram and cumulative density function of the quarantine period

Figure 4.9 displays the survival probability curves of users in terms of time spent in the quarantine environment for the first three quarantine events and the rest. As demonstrated in Figure 4.9, end users spent more time in quarantine during their first time than the second time. This might be due to being unfamiliar with the environment or with the process to clean up the infection.

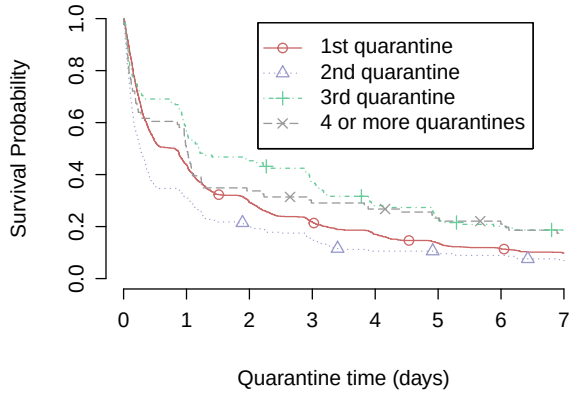


Figure 4.9: Survival probabilities over different quarantine events

To further investigate, Table 4.6 shows the median time spent in quarantine during the first three quarantine events. We compare them across the different release mechanisms and cleanup outcomes. End users that managed to remove the infection, stayed longer in the walled garden than those who remained infected, regardless of which release mechanism was used. Take a look at the median time of the assisted end users in the first

Table 4.6: Summary statistics on the time to cleanup for self released and ISP assisted released mechanisms

Quarantine Event Number	Status	Total # Users	Median quarantined time (hours)		
			Cleaned, not seen again	Cleaned, later reinfected	Still Infected
1	Self release	805	11.16	10.69	10.24
	Assisted	361	24.25	4.90	6.82
2	Self release	195	8.07	7.04	3.74
	Assisted	102	24.87	25.28	3.30
3	Self release	17	3.15	11.42	3.29
	Assisted	114	69.72	49.60	22.51

quarantine event, for example. Those who managed to clean up spent 24 hours in quarantine, while users who remained infected took just around 7 hours. In the self-release group, successful cleanup also took longer, though for the first quarantine event, the difference is surprisingly small with the group that remains infected or got re-infected at a later stage (roughly 11 versus 10 hours).

During the second and third quarantine event, the differences become more pronounced. Longer time spent in quarantine is now clearly related to cleanup success. Users who remain infected spend about half as long in quarantine as the other two groups. It seems a certain group of users is becoming habituated to the walled garden notification and environment. They self-release very quickly and it seems unlikely that they made a serious attempt to perform cleanup.

It is important to note, though, that the self-releasing users that do succeed in cleaning up also leave the walled garden faster over successive quarantine events. The median time drops from 11 hours during the first quarantine to 8 hours (second quarantine) and then to just 3 hours (third quarantine). In other words, it seems there is not just habituation going on, but also actual positive learning effects in terms of how to perform cleanup and navigate the release from the walled garden.

4.6 End user reactions

To get a better sense of the actual experience of the end users, we qualitatively analyzed the communication of the quarantined users with the abuse and support staff at the ISP. Each communication channel was used for different of reasons. Generally, emails were sent to inform abuse desk employees about the cleanup efforts and possible causes of the

infection. Interaction with the support staff, on the other hand, were more often asking for more information about the quarantine and how to resolve the situation. The content of the submitted walled garden forms often contained more specific information on the cleanup actions taken by the quarantined users. For instance, some users pasted the output of the antivirus scans in these forms to prove that the infection was no longer present.

First, we manually analyzed a sample of 200 walled garden forms, 200 help desk logs and 50 emails to the abuse desk. We saw five recurring themes that speak to the user experiences of the walled garden: (i) asking for additional help to resolve the infection and leave the walled garden; (ii) requesting a paid technician to visit the user; (iii) expressing distrust of the walled garden notification; (iv) complaining about the disruption of service; and (v) threatening to terminate the contract with the ISP. To get a sense of how many users were associated with these types of communications, we collected keywords from the manual analysis of the sample and then searched the full communication data for their presence. Table 4.7 shows the number of unique users associated with each topic. For 51% of the users who communicated with the ISP, their messages did not fit any of these topics and we categorized them as 'Miscellaneous'.

4.6.1 Requesting additional help

Almost 27% of the users at some point contacted the ISP to ask for additional help to cleanup the infection. The users wanted to solve the problem, but they were unable to understand the notification or to follow the steps towards quarantine release. The type of help that is requested varies widely. Some of this is driven by differences in the type of infection and the operating system of the user. Cleanup software and materials provided in the notification content would not work on all OS types, OS versions and patch levels. Some customers in our study downloaded the requested software to remove the infection, only to find out that it would not install correctly. Some users could not download the software at all from the links provided by the ISP. In those cases, they requested to be released from the quarantine environment so that they could download additional software.

One of the malware families was Mirai – the infamous botnet made up of Internet-of-Things devices. Not surprisingly, users with these infections asked for help in identifying which of their many devices was the problem and how to then secure it from future infections. Not to put too fine a point on it, but from a usability perspective the cleanup of compromised IoT is a world of pain for which we have very little practical guidance. In these cases, ISP staff would ask users additional questions about what devices they had connected to their home network. Based on the replies, staff would try to identify the offending device and more specific cleanup actions. In one case, after contacting the ISP, a user disconnected his IP camera from the network so as to prevent future infections and quarantine events, while the actual problem later turned out to be a DVR. The user ended up getting infected and quarantined again.

4.6.2 Requesting a paid technician

About 7% of the users in our study were not capable of removing the infection by themselves and requested the ISP to send a paid technician to their home. In a handful of cases, end users mentioned taking their computer to technicians at local computer repair shops. The ISP's technicians are typically people who also have a background in abuse handling. Some of the communications we analyzed were from these technicians themselves who contacted their colleagues at the ISP abuse department from the customer premises and provided detailed information about their cleanup actions. This way, the abuse desk employees got the required proof of cleanup and could release the connection from the walled garden. Interestingly, in a few rare cases, we found that the paid technician could not actually find the infection. They then referred the end users back to abuse desk employees to communicate the occurrence of a false positive. Unfortunately, as a result of this process, users remained in the walled garden environment longer.

4.6.3 Distrust of the notification

Around 2% of the users contacted ISP employees to confirm the veracity of the email and walled garden notifications. They did not expect that their ISP would notify them about an infection and were worried that this could be a phishing attack to install ransomware or steal personal information. Users mainly contacted help desk employees to confirm the veracity of the notifications. One user replied directly to the notification email, i.e., using the very channel that he did not trust, and voiced his concerns this way to the recipient at the abuse department.

4.6.4 Complaints over disruption of service

Placing a customer in a walled garden environment is a strong incentive for end users to clean up, but also an intrusive measure. During in our study period, around 10% of the users complained in some shape or form. Some reported that their business was disrupted due to having no Internet to work with. Usually, these turned out to be users that run small businesses over their consumer broadband connection: shops, restaurants and even a small medical clinic. They claimed that they could not provide services to their customers and, as a consequence, lost customers. Some mentioned, for example, that the payment terminals did not work and so their customers could not complete their purchases. In two cases, the owner of the shop stated he had to close the shop until the problem was fixed. Several of these users provided a calculation of the monetary loss they suffered and demanded a reimbursement from the ISP.

4.6.5 Threats to terminate the contract

Around 3% of the users were so unhappy about their connection getting quarantined that they threatened to terminate their subscription and move to one of the ISP's competitors. Some of the users pointed to the losses they had incurred, others to the fact that they had to pay for the subscription even though they no longer were provided with Internet access. Also several users threaten to leave the ISP because the user could not, even with their best effort, identify and remove the infection. These users were quarantined multiple times and they spent quite a bit of time in the walled garden environment.

Table 4.7: User issues raised in communication with ISP

Topics	**# of users**
Request additional help	323 (27 %)
Request paid technician	80 (7 %)
Distrust of the notification	19 (2 %)
Complain over disruption of service	126 (10 %)
Threaten to terminate the contract	39 (3 %)
Miscellaneous	621 (51 %)

4.7 Ethical Considerations

Access to data about the user's experience upon abuse notifications is extremely limited and cooperation with an ISP is essential to enable otherwise impossible research. For this study we leverage secondary data that was originally collected by an ISP for business purposes. This data was pre-processed by a coauthor of this paper while working for this ISP and with the consent of the ISP's abuse desk manager. Moreover, the data was processed on the ISP premise and within the ISP privacy policies.

Unavoidably, the processed dataset was not fully anonymized as the high dimensionality of the data did not allow for a robust anonymization, i.e., the anonymization would have led to an unacceptable level of data loss. To ensure confidentiality, the raw dataset was stored in a secure server to which only authorized users could access. Moreover, the data was analyzed while preserving the privacy of the ISP's customers and ensuring that it is not possible to identify them from any of our results. Both the processed and anonymized data were removed after the publication. The original data remains in the ISP systems, allowing for replication if needed.

4.8 Limitations

We underline four limitations relevant to the findings of our study. First, we based our study on a single ISP with a relatively strict implementation of the walled garden notification system. The generalizability of our results to other implementations and ISPs is a matter for further studies. Second, our study uses data collected as part of the operational process of the ISP. As such, the study lacks an experimental design and a control group. This means we cannot compare the effectiveness of the walled garden notification to the cleanup rate of a mere email or no notification whatsoever. Third, our dataset on infections is limited to what has been reported in the Shadowserver feeds. As a result of this, we lack visibility into notifications triggered by other feeds and infections that are not reported by Shadowserver. This makes our coverage of malware infections biased towards those that are sinkholed and reported by Shadowserver. Malware that has escaped these defender efforts might also be harder to clean. Fourth, the cleanup outcomes are also based on the Shadowserver feeds. It is possible that an infection might not show up in the Shadowserver feeds right away. This is partly driven by user behavior, such as temporarily turning off the infected device or disconnecting it from the Internet, and partly by other factors, such as the properties of the malware families. Some are less aggressive in terms of scanning for victims or contacting the command-and-control server for commands. This absence in the feed may cause us to overestimate the cleanup rate. For this reason we chose a conservative time frame. We only counted a machine as cleaned up if we did not see it for 30 days after release from the walled garden.

4.9 Conclusion

In this study, we explored the effectiveness of walled garden notifications and quarantining in terms of helping users in residential networks to perform malware cleanup. Based on data on $1,736$ quarantining actions involving $1,208$ unique users, collected from April 2017-October 2017 by a medium-sized European ISP, we found that roughly half to three quarters of the quarantined users had managed to clean their machine. There is no clear point of reference for this success rate. When we look at prior work on abuse and vulnerability notifications, it seems to be quite high. Most of those studies find rates well below 50%. That being said, comparison is difficult as the typical recipient of those notifications is a server admin or webmaster, not a home user.

Most users are quarantined only once, so the effort of cleanup kept them clean for months, if not longer. Perhaps the quarantine experience made users adapt their online behavior or improve their system's security defaults, like automatic patching and the installation of antivirus tools. This suggests there may also be long-term benefits to quarantining, beyond mitigating the immediate threat posed by the infection.

Users could self-release easily and quickly for the first two quarantine events in a month. Remarkably, this easy way out does not incite lax security behavior. Cleanup rates

are either as high, or just a bit lower, than users who have to submit proof of cleanup to the provider and wait for the abuse staff to release them. We see a bit of evidence for habituation among a small group of users who learn how to release themselves from quarantine, rather than clean the infection. We also saw evidence, however, of a positive learning effect: successful cleanup also became faster for users going through successive quarantining events.

All in all, we found substantial support for the effectiveness of this best practice for ISPs in the fight against botnets. Since effectiveness of the other recommended best practices has been questioned, this suggests more ISPs should be considering to adopt a walled-garden solution. In light of the rising problem with IoT malware, this might become a critical line of defense. That being said, IoT malware remediation methods will differ from traditional cleanup strategies and, thus, walled garden implementations will have to be revisited to accommodate the cleanup requirements for IoT malware.

On the downside: setting up and maintaining a walled garden environment is a significant investment for an ISP. Furthermore, providing support to users in their attempts to clean up also imposes a significant cost. Around one out of four quarantined users posed a question for help to a staff member. These costs could perhaps be reduced by allowing self-release more broadly, since it seems to be more or less equally effective as the more labor-intensive form of provider-assisted release. Some of this assistance might provide a business opportunity, as we found that around 7% of the quarantined users asked for a paid technician.

A fraction of the users, around 10% of them, voiced complaints over the disruption. Around 3% even threatened to terminate the contract. We do not know how many users actually terminated their subscription, but the threat alone might, unfortunately, be enough to scare off some ISPs from investing in a walled garden. In competitive broadband markets with high penetration rates, customer acquisition is very expensive. In these situations, a prisoner dilemma might appear as not having a walled garden might be a competitive advantage. This could push ISPs to not deploy it, even though it is effective. On the other hand, if all ISPs adopted it simultaneously, it would generate collective benefits, though these would not necessarily flow back to the ISP, except through lower customer churn rates.

We did notice that the group which seemed the most negative about the quarantining actions were small businesses operating on a consumer broadband connection. ISPs could prevent them from being affected in the future by providing an easy transition to a comparatively-priced business subscription, which would take them out of the consumer market – and thus keep them away from the walled garden. This would reduce the push-back over time and allow the walled garden to do what it does best: protecting home users from further damage caused by their infection, and protecting the rest of the Internet from the infected home user.

CHAPTER 5

Evaluating ISP-made vulnerability notifications

Mechanisms for large-scale vulnerability notifications in chapter 3 and previous research have been confronted with disappointing remediation rates[44, 45]. It has proven difficult to reach the relevant party and, once reached, to incentivize them to act. We present the first empirical study of a potentially more effective mechanism: quarantining the vulnerable resource until it is remediated. We have measured the remediation rates achieved by a medium-sized ISP. In this chapter, we assess the effectiveness of quarantining by comparing remediation with two other groups: one group which was notified but not quarantined and another group where no action was taken. The chapter also investigates issues raised by recipients of ISP-made vulnerability notifications by looking at the communication between ISP and notified end users.

5.1 Introduction

Our ability to undertake large-scale vulnerability discovery has grown immensely, providing a wealth of data on vulnerable resources to help those responsible for the affected resources. Notification and remediation, however, has proven to be much harder. Randomized controlled experiments with different notification mechanisms have found remediation rates that typically range from modest to abysmal. These low rates persisted across disclosures via email, national CSIRTs (Computer Security Incident Response Teams), social networks and even phone calls [43, 92, 45, 44].

There are varying explanations for the disappointing remediation rates. Most experiments used email. Because it scales reasonably well, this is still the dominant channel for notifications. Reachability has proven to be a key problem, however. Notifications are sent to addresses that are RFC-specified or harvested from WHOIS records. Delivery is severely hampered by non-existing email addresses and poorly-configured spam filters. When messages are actually received and read, there is often no follow-up action. These problems are not specific to email. Even more manual methods for notifications, such as postal mail or phone calls, have the same issue [45]. The lack of follow-up actions points to problems

with trust, technical competency and lack of incentives for remediation.

One would expect that the incentive problem would be even worse for vulnerabilities that threaten third parties rather than the party responsible for the vulnerable resource. Think of NTP servers that can be abused in UDP-based amplification DDoS attacks against any target on the Internet. They are rarely, if ever, used against the party responsible for the vulnerable server itself. Remarkably, though, a 2013 campaign of researchers and the security community managed to reduce the number of vulnerable NTP amplifiers by more than 92% in three months [42].

This stand-out success has been difficult to interpret, partly because it was not a randomized controlled experiment. A high-profile campaign that did use an experimental design, was Heartbleed [41]. It also found a relatively high overall remediation rate of around 60% over the course of a month. While these examples provide inspiring counterpoints to the studies with disappointing remediation rates, these high-profile campaigns do not seem suitable templates for large-scale vulnerability notifications.

All in all, prior work on notifications has observed an alarming and increasing discrepancy between the community's ability to gather vulnerability data and its ability to make this information useful for preventing future abuse. In this paper, we empirically explore the effectiveness of an alternative mechanism for vulnerability notification and remediation: quarantining the vulnerable resource in a so-called walled garden environment. We compare this to the current default approach: email notifications. Walled gardens tackle both challenges identified in previous studies. First, it provides a much more robust mechanism to notify the responsible party, as Internet access is restricted and a landing page informs the party responsible for the vulnerable device of the reason why the connection is quarantined. In other words, it is almost impossible to overlook the notification. Second, the mechanism increases the incentive to remediate, as release from the walled garden is conditional on remediation. Prior studies has found that quarantining was effective in cleaning malware infections in ISP networks[93, 94]. Its effectiveness in remediating vulnerable resources, however, has never been studied before.

We study a walled garden implementation for vulnerable resources at a medium-sized Internet Service Provider (ISP). We measured remediation rates for $1,688$ retail customers with servers running open DNS resolvers or Multicast DNS services, which can be abused in amplification DDoS attacks. We assess the effectiveness of quarantining by comparing remediation with two other groups: one group which was only notified by an email but not quarantined and another group where no action was taken.

In short, we make the following contributions:

- We present the first empirical study of the remediation effectiveness of quarantining vulnerable resources. Even though customers can self-release from the quarantine environment without actually remediating the problem, we find very high remediation rates of around 87%. Of those who received only the email notification, around 75% remediated.
- We find a remarkably high remediation rate in the control group: around half of all

customers remediate. This high rate reflects actual remediation actions, but also the fact that a significant portion of the observations of vulnerable devices are transient. These observations have typically been omitted from prior studies, which might explain the low remediation rates reported in those papers. This might reflect selection bias.

- We analyze communications between notified customers and the ISP to assess challenges in remediation. We find out that 16% of the notified users were unwilling to remediate because they did not want to change the way they use their device. Around 11% of the notified users complained about the disruptiveness of being quarantined in the walled garden.

Notwithstanding the potential advantages of walled garden solutions, we want to emphasize that quarantining vulnerable resources is not a silver bullet. Quarantining by network operators is only feasible under certain scenarios. There are also downsides in terms of cost and customer pushback. We will discuss these in the course of the paper. We do argue, however, that there is an urgent need to find more effective notification and remediation mechanisms. This puts a premium on examining solutions for which no prior empirical studies exist.

This chapter is structured as follows. Section 5.2 explains the unique natural experiment that was inadvertently conducted by a European ISP. Section 5.3 describes the data collection mechanism. Section 5.4 evaluates the effectiveness of walled garden and email notification mechanisms compared to natural remediation. Section 5.5 presents key insights gathered from communications. Section 5.6 evaluates prior work and explains how this is related to ours. We outline ethical considerations and limitations of the study in Section 5.7 and 5.8 and conclude the study in Section 5.9.

5.2 Vulnerability notification experiment

For this study, we collaborated closely with a European ISP which operates in various markets. Here, we will focus on its retail broadband services, which have around 2 million customers. A few years ago, the ISP implemented its first version of a walled garden solution to deal with malware infections among its retail customers. More recently, the ISP started allocating spare capacity in the walled garden environment to undertake notification and remediation for users with devices that are vulnerable to UDP-based amplification attacks, as identified in Shadowserver scans for such amplification factors. The ISP only does these notifications on a fixed day each week. This setup provides a natural experiment, as the assignment of customers to one of three groups (quarantine, email, no action) is more or less random.

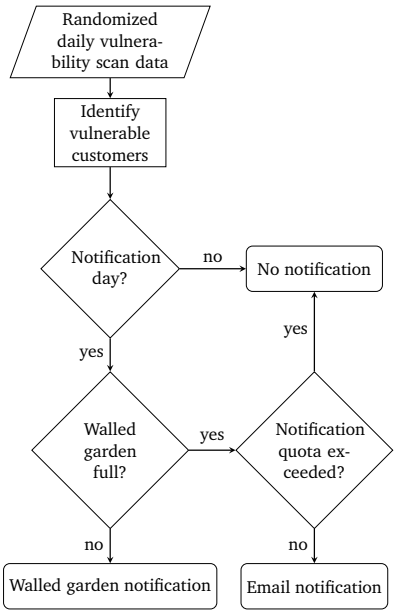


Figure 5.1: Vulnerability notification flowchart

5.2.1 Walled garden notifications

In the early days of the Internet, the concept of a "walled garden" referred to a closed environment that restricted the content and services that users could access. Nowadays, a walled garden primarily refers to a security best practice in botnet mitigation [78], as described in RFC6561 [69]. It is a method to notify affected users about a security problem and quarantine their connection to prevent the infected machine from being abused by miscreants.

The ISP with which we collaborated has adopted a so-called strict implementation of a walled garden. This means that the quarantine network redirects all web browsing activity to a landing page, except for a small set of white-listed sites. The landing page explains the problem and provides guidance on resolving it (see Appendix D.1). The advantage of this notification mechanism, compared to email, postal and phone notifications, is that it is much less likely to be overlooked or ignored. At the same time that the connection is quarantined, the ISP sends an email to the customers with the same information as the landing page. Thus, users don't need to be at their home to understand that their Internet connection has been quarantined by the ISP.

There are three ways the customer can get out of the walled garden. First and foremost, customers can release themselves from the quarantine environment via a button underneath a form for reporting on what action was taken. The self-release option is revoked after two subsequent quarantine events in the same month, to avoid customers using this route to restore their connection without making an effort at remediation. The second way out is when the ISP's abuse staff releases the customer's connection. Customers might end up in assisted release because they no longer have the self-release option or because they have contacted the ISP for help. Quarantined customers can contact abuse desk members via email and a walled-garden form. The third way of being released is when the expiration date passes. After 30 days, a customer is automatically released, even if they have not contacted the ISP.

5.2.2 Email notifications

The walled garden has a limited capacity. When all slots are taken, but the ISP still wants to notify and remediate, it can send an email notification to the mail address that it has on record as the primary contact for that customer. For some customers, the ISP's mail service is the primary contact point. For other users, it does not have full visibility into the delivery success of the message. That being said, these are email addresses that were supplied by the customers themselves, so the odds of success are a lot higher than mailing RFC-specified addresses or generic WHOIS contact points. The message contains the same information as the walled garden's landing page, plus an email address to contact in case of questions or problems while remediating the vulnerability.

5.2.3 Notification process and assignment mechanisms

On a daily basis, the ISP receives vulnerability scan data from third parties, most notably Shadowserver, specifying a list of vulnerable IP addresses in the network. IP addresses show up in the daily vulnerability scan data in a random order. Because of time constraints of the abuse department, the ISP notifies owners of these resources only once per week, with different vulnerabilities being assigned different weekdays. For mDNS and open resolver notifications are made every Thursday, using the IP addresses from Wednesday's reports. This arbitrary policy and randomized list of IP addresses in the vulnerability scan data create a natural experiment: a *de facto* random assignment to being notified or not notified, assuming that there is no systematic difference between customers that show up in Wednesday's reports versus the reports from, say, Tuesday or Thursday.

The next step contains a random assignment between the two treatment conditions: walled garden and email-only notifications. The ISP's walled garden can fit up to 100 customers at any time. Many of the slots are taken for higher priority issues, such as malware infections. The remaining slots are dedicated to a random batch of customers selected from Wednesday's Shadowserver report, without any prior inspection of the IP

addresses in the report. When full capacity is reached, the remaining customers are notified via email, until also for that treatment a quota is reached. The quota is a bit fuzzy and depends on the available resources (e.g., abuse department staff, number of open tickets, etc.). If the walled garden capacity and the email notification quota are both exceeded, then the remaining vulnerable customers are not notified. Figure 5.1 shows a flowchart of the treatment assignment process. A direct consequence of only notifying once per week is a higher amount of vulnerable customers that do not receive the treatment compared to the ones that are not notified. This imbalance will increase the power of the natural experiment even though the groups are asymmetric in size.

Given this notification process, the treatment assignment is independent of the characteristics of the vulnerable population. When after a notification a customer machine shows up again in the Shadowserver reports, then the assignment process may result in a subsequent treatment. Also, customers may have other vulnerabilities on their machine and they may also receive notifications for these issues via a different procedure, delivered on other weekdays. This may also impact the remediation of mDNS and Open resolver vulnerabilities. In our statistical analysis, we use an instrumental variable to account for this effect on the vulnerability remediation.

5.2.4 Other walled garden notifications

As this experiment was conducted in a real-world setting, we also had to take into account that the ISP sent out notifications for other security and vulnerability issues that were not part of the experiment. Checking the ISP logs, we found out that 231 users in our study did in fact receive another walled garden notification (16% of the users in the control group and 8% of the users in the treatment groups, see Table 5.3). The bulk of these notifications (95%) were for NetBIOS. Like mDNS and Open resolver, NetBIOS can be abused in amplification attacks. Rather than removing these users from the study, we decided to keep them in and use this opportunity to study the impact of other notification processes. Most real-world randomized controlled notification experiments are likely impacted by unobserved 'parallel' notifications processes. In those cases, the researchers typically have no data on this. In our study, we did have the data, so we were in a position to identify just how these 'other notifications' impacted the results.

5.3 Data Collection

To assess the notification and remediation success of the walled garden solution, we correlate three different datasets collected by the ISP: (*i*) Daily scan results on the presence of vulnerable amplifiers in the ISP's network, provided to the ISP by the Shadowserver Foundation; (*ii*) ISP logs that capture the details of all walled garden or email notifications; and (*iii*) abuse desk emails and walled garden contact forms that capture the communication

flows between abuse department and customers. All in all, our data covers $1,688$ unique customers who were seen to operate vulnerable devices in the ISP's consumer network between September 26th, 2017 and December 31th, 2017.

5.3.1 Vulnerability feeds

The ISP receives a daily report on vulnerable devices from the Shadowserver Foundation. These daily feeds not only identify new vulnerable devices, but also allow us to track if a device is remediated after a quarantine event or an email notification. We selected two types of vulnerabilities based on:

- *mDNS reports*: Multicast DNS (mDNS) reports provide the results from scans for publicly accessible devices that have the mDNS service accessible and answering queries. In the period of our study, a total of $1,575$ customers were found with vulnerable devices.
- *Open resolver reports:* Shadowserver open resolver reports contain information about publicly-available recursive DNS servers. Throughout our study period, we identified 113 customers with such a vulnerable device.

Table 5.1: Vulnerable hosts and percentage notified

	mDNS	open resolver
# vuln. hosts	$1,575$	113
% notified	474 (30.09%)	22 (19.46%)

A daily breakdown of the number of customers reported in the feeds is shown in Figure 5.2. Between October 26 and November 6, 2017, the ISP did not receive any reports from Shadowserver due to server maintenance. No notifications are made during this period. Table 5.1 shows what fraction of the affected customers were notified via email or the walled garden.

The ISP does no prior filtering or inspection of the IP addresses in the Shadowserver reports before assigning treatments. This means that notifications are made irrespective of how often the IP address or customer has been seen in the reports. This is different from how most vulnerability notification experiments have been designed, where notifications are typically restricted to devices that are consistently seen over a certain period to avoid including false positive or more transient issues (e.g., [43]).

The downside of our approach, or rather the ISP's approach, is that we likely overestimate the remediation rate, as some of these devices that disappear from the reports reflect

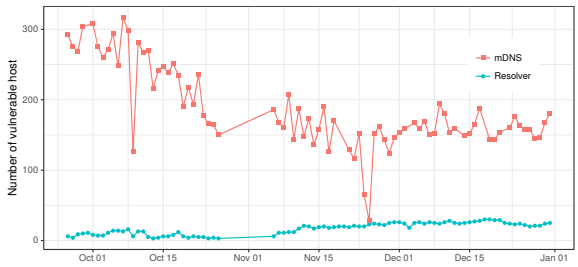


Figure 5.2: Daily number of vulnerable hosts during the observation period

not actually remediation but transient issues. The upside is that we do not introduce selection bias. Including only those devices that are consistently seen as vulnerable over a longer period is likely to restrict the study to a non-representative subset of all vulnerable devices and their owners. Home users whose devices occasionally power off or go into standby might get excluded, for example. In other words, there is a trade-off between selection bias and overestimating the remediation rate. We decided to tolerate the latter rather than the former and to not exclude any cases that were part of Shadowserver reports and the ISP's process.

5.3.2 Notification logs

During our study period, 350 walled garden notifications were made to 327 users and 322 email-only notifications were sent to 249 users. Some users were notified more than once, sometimes with different notification types. Of all 1,688 customers in the Shadowserver reports on mDNS and open resolver, 279 also received a walled-garden notification and 3 an email-only notification for another vulnerability during the study period. For each of these notifications, we gathered (i) notification time; (ii) notification type; (iii) number of notifications made; and (iv) reason for the notification. Additionally, for walled garden notifications, we collected (i) quarantine start date; (ii) quarantine release mechanism; and (iii) quarantine removal timestamp.

5.3.3 Abuse desk logs

Notified customers can respond to the notifications via emails sent to the abuse team or, when in quarantine, via a contact form on the landing page. To better understand how users reacted to the notification and quarantine events, we gathered 564 emails from 261 users and 324 walled garden forms from 232 users.

5.4 Results

In this section we evaluate the effectiveness of the notification mechanism by investigating the percentage of users that remediated in each of the following three groups: (*i*) notified and quarantined (walled garden), (*ii*) notified but not quarantined (email), and (*iii*) no action. We measure remediation via the daily reports provided by Shadowserver. There are various reasons other than remediation that might make a device disappear temporarily from the feeds, such as a temporary shutdown of the device or a disruption in the network. To conservatively estimate remediation, we check whether a vulnerable device shows up in the Shadowserver reports after the notification period, between January 1 - 31, 2018. If we do not see the device in the reports for the whole month, we assume it is remediated. This approach means we do not estimate remediation speed.

5.4.1 Measuring the impact of notifications

We first study the difference in remediation rates among the three groups: walled garden, email-only, mixed notifications and no notification (control). For this comparison, we investigate what portion of users in control group received ISP notifications for other security problems in the same observation period. This turned out to be the case for 192 users in the control group. In the same observational period, 95% of the other notifications were made for publicly-accessible devices with vulnerable NetBIOS services. Like mDNS and Open resolver, NetBIOS can be abused in amplification attacks.

While investigating the rates of remediation for the control group, we had to take into account the presence of other ISP walled garden notifications. For this reason, we divided the control group into 2 groups: (i) users who received other security notifications from the ISP; and (ii) users who received no notifications whatsoever. Table 5.2 shows the remediation rate for users in the control group who received other notifications compared to users did not receive any notifications: 96% versus 53%, respectively. A plausible explanation for this high impact of other notifications is that the typical remediation actions for NetBIOS also impact the mDNS and Open resolver vulnerabilities, e.g., disabling the DMZ or taking the device offline altogether. Table 5.3 shows that a small subset of users in the treatment groups also received other walled garden notifications. These show high remediation rates as well, but the difference is more modest compared to the other users in these groups. Note that later in this section (See section 5.4.5), we will present a logistic regression model that systematically controls for the impact of other notifications while estimating remediation rates for the different experimental groups.

Table 5.2 also shows that notifications for the actual vulnerability have a clear impact on its remediation. Around 87% of users in the walled-garden group remediated compared to 75% of users in the email-only group. Moreover, users that received both email and walled garden notifications on different days remediated around 81%. While the walled garden is clearly highly effective, the control group remediation rate is also surprisingly

Table 5.2: Summary statistics on the percentage of remediation according to the treatment groups and control group

Type of Notification	Status	Total # Users	Remediation Rate
Only walled garden notifications	mDNS	225	194 (86.2%)
	Open resolver	22	20 (90.9%)
	Total	247	214 (86.6%)
Only email notifications	mDNS	169	127 (75.1%)
	Open resolver	-	-
	Total	169	127 (75.1%)
Mixed notifications	mDNS	80	65 (81.25%)
	Open resolver	-	-
	Total	80	65 (81.25%)
Control other walled garden notifications	mDNS	181	175 (96.6%)
	Open resolver	11	10 (90.9%)
	Total	192	185 (96.3%)
Control no notifications	mDNS	920	484 (52.6)
	Open resolver	80	48 (60.0%)
	Total	1000	532 (53.2%)

high: around 53% for the ones without any notifications and 96% for the ones that received other notifications. We will revisit this issue in the next subsection. Overall, remediation rates are high. This stands in stark contrast to most prior studies and is in the same range as the two high-profile cases of NTP amplifiers [42] and servers with the Heartbleed vulnerability [41].

5.4.2 Natural remediation

How can we make sense of the remarkably high remediation rates in the control group, even when we exclude the group who was notified for a different security issue? We consider two potential explanations: (i) transient events; and (ii) DHCP churn effects. Below, we explore the possible influence of each factor.

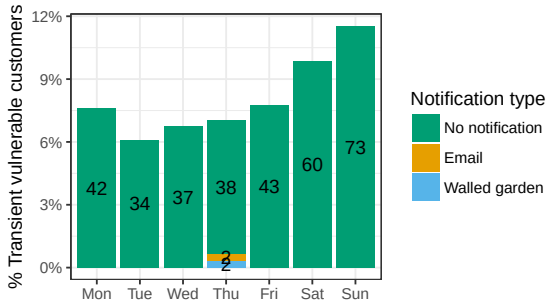


Figure 5.3: Percentage of transient vs. non-transient vulnerable customers per weekday

We first investigated role of transient events, as significant portion of the vulnerable devices reported by Shadowserver capture transient events. As discussed in 5.3.1, we did not exclude any vulnerable devices from our study to avoid selection bias. This means that remediation rates are likely overestimated by counting transient events as remediation.

This seems to impact the control group more than other groups. As figure 5.3 shows, users who did not receive any notifications have a larger fraction of observations that are seen once or twice in comparison to the notified users. In total, there were 331 transient vulnerable customers of which only 4 received a notification. This is mainly due to the notification process in itself, as there is a larger fraction of transient events during the non-notification days. This is specially prominent during the weekend when the proportion of transient events increases from 20 to 40% compared to working days. This might be due to typical use cases for mDNS, namely music sharing and video streaming between devices on a home network during the weekend. As devices move from their local home networks to other networks, such as a friend's house, their mDNS functionality temporarily appears in other networks[95]. In this short period, they then appear in the vulnerability scan data. Figure 5.3 shows this pattern by visualizing the percentage of transient events, calculated as the ratio of vulnerable customers that are only reported once divided by the total amount of reported vulnerable customers per weekday.

While almost 30% of the users that did not receive notifications were seen once, only less than one percent of the notified group was seen only once. This shows that it is more likely to overestimate remediation rates of non-notified users than the ones that receive notifications. If all devices that are seen once are transient vulnerabilities, then this would already explain around half of the remediation rate of the control group.

Figure 5.4 evidences a strong correlation between the endogenous explanatory variables (i.e., the notification type) and the frequency at which a vulnerable resource is re-

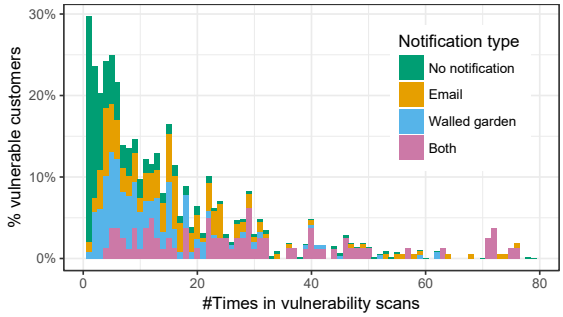


Figure 5.4: Distribution of vulnerable customers appearance in the feeds

Table 5.3: Remediation rates for users in different groups who also received other notifications

	Other notifications	
	#	remediation rate
Only walled garden	26	24 (92.3%)
Only email	11	10 (90.9%)
Mixed notifications	2	2 (100%)
Control	192	185 (96.3%)

ported. Hence this frequency can be used as an instrumental variable for the consistent estimation of the remediation rates. Note that this does not invalidate the inherent randomized assignment of the notification process as this frequency is not a characteristic of the vulnerable customers, and there is no reason to believe that the characteristics of the population that present a transient vulnerability are different from those with long-lived vulnerabilities. We leverage the amount of times a vulnerable customer appears in the reports as an instrument to account for its impact on the treatment (notifications) which in turn impacts the remediation occurrence.

Lastly, we looked at the impact of DHCP churn on remediation rates. The ISP assigns dynamic IP address with very long DHCP lease times, typically a year. This means that a certain portion of customers has been assigned a new lease, over the course of the measurement period. This impacts our measurements for the control group differently that those for the treatment groups. The ISP's abuse department stores the IP addresses of the notified customers and we can track whether a customer has been assigned different addresses over the period of the study. The ISP does not store the IP addresses for the customers who were

not notified. We had to look these up ourselves at the end of the measurement period. In other words, we could not control for churn in the control group that did not receive any security notification. As a result of this, remediation rates for 192 subscribers in control group that received security notifications for different issues were not influenced by DHCP churn. On the other hand, remediation rates for the rest of the control group might have been influenced by DHCP churn. When a vulnerable device changes its IP address, we will see the device at the old address as remediated. This means that in the control group, we will overestimate the remediation rate because of DHCP churn.

All in all, while we have no definitive explanation, these two factors help understand why the remediation rate of the control group might have been so high.Transient observations affect the remediation rates in all groups, but the control group most of all. The rate might have been further impacted by DHCP churn. This means that in reality, the difference in remediation rates between control and treatments is likely to be larger than we reported. Our conclusion that the treatments have a significantly higher impact compared to the control group is, therefore, not affected by these issues.

5.4.3 Release mechanism

We have seen that remediation rate for the walled garden group was higher than for the email-only group. The core difference between those two treatments is the incentive they provide. An email can be easily ignored, while the walled garden more forcefully compels users to act. That being said, customers can self-release from quarantine with the push of a button. In other words, if they are unable or unwilling to remediate, the walled garden does not stop them from leaving – at least not for the first two quarantine events within a month.

To see if self-release is associated with lower remediation rates, we take a closer look at the results for the different release mechanisms for users that were quarantined once: self-release, ISP-assisted release or release because the maximum quarantine period of 30 days expired.

From Table 5.4, we can observe that out of 236 total users, 156 (66%) used the self-release option and 86% of them remediated the vulnerability while they were in the quarantine. This is only marginally lower than the 90% remediation rate for the 79 (33%) users who contacted ISP staff for assisted release. Just one user did not use either one of these options and his or her device was also remediated. All and all, it seems self-release did not negatively affect remediation success. The incentive mechanism worked well without being overly stringent and allowing users a speedy release and restoration of their Internet connection.

Our results show a higher remediation rate than observed by a prior study on quarantining ISP customers with a malware infection [94]. It found that quarantining incentivized 69% of 1,208 infected end-users to cleanup after the first event. The difference might be due to the fact that the vulnerabilities we studied are more transient in nature than mal-

ware infections. In both studies, assisted users showed slightly better remediation rates than the users who self-released from the walled garden.

Table 5.4: Release types and remediation

Status	1st Quarantine Event	
	Total # users	Remediation rate after Q
Self release	156	134 (85.8%)
Assisted	79	71 (89.8%)
Expired	1	1 (100%)
Total	236	206 (87.2%)

5.4.4 Measuring the impact of multiple notifications

We now take a closer look at the users who received more than one notification after they did not manage to remediate the vulnerability. Table 5.5 reports the remediation rates for these users. We separate users for whom the subsequent treatment were the same from those who received a mix of treatments. As table 5.5 demonstrates, the pattern is consistent with our earlier findings: the email-only treatment has a lower remediation success than the walled garden. Remarkably, the mixed treatments have an even higher remediation rate. We have no explanation for this result. One speculation is that it reflects how the user interprets the walled garden notification. If that treatment came first, then the subsequent email-only notification may serve as a warning that the connection might be disrupted again if the user does not act. If the email-only treatment comes first, then the subsequent walled garden action might be seen as an escalation process, compelling the user to act before further consequences are imposed.

Table 5.5: Remediation after multiple notifications

	2 notifications		3 or more notifications	
	#	remediation rate	#	remediation rate
Only walled garden	11	8 (72.7%)	-	-
Only email	25	17 (68.0%)	8	6 (75.0%)
Mixed treatment	47	41 (87.2%)	33	24 (72.7%)

5.4.5 Modeling remediation occurrence

In this section we further investigate the direction and magnitude of different factors on remediation success. We investigate several observable characteristics of notifications. We use a multivariate logistic regression model that takes five explanatory (independent) variables as input:

- x_1: **Type of notification**: Categorical variable that represents the type of notification used. In our experiment we had 2 different types of notifications: (i) email and (ii) walled garden. This variable captures if one or both notification types are sent.
  - **Only email notification**: This represents users that receive only email notifications.
  - **Only walled garden notification**: This represents users notified through only walled garden notifications.
  - **Mixed notifications**: This represents users that received both email and walled garden notifications, but on different notification days.
- x_2: **Number of walled garden notifications**: Total number of walled garden notifications made per vulnerable user.
- x_3: **Number of other walled garden notifications**: Number of notifications made to a user to remediate other types of vulnerabilities. This variable captures if a user in one of the treatment of control groups received other notifications and, if so, how many. This variable allows us to distinguish two subgroups in the control group: 1,000 users who did not receive any notifications versus the 192 users who did receive a walled garden notification for another security issue (see Table 5.2).
- x_4: **Number of email notifications**: Total number of email notifications made per vulnerable user.
- x_5: **Type of Vulnerability**: Categorical variable that shows the type of vulnerability.

These explanatory variables are included in a multivariate logistic regression model to estimate the probability of remediation occurrence. The binary logistic regression equation is explained as:

$$logit(\pi_b) = log\left[\frac{\pi_b}{1-\pi_b}\right], \tag{5.1}$$

where π_b is the probability of remediation within the range [0, 1] and is estimated as:

$$\pi_b = \frac{exp(\beta_0 + \sum_i \beta_i x_i)}{1 + exp(\beta_0 + \sum_i \beta_i x_i)}, \tag{5.2}$$

where x_i ($i = 1, \ldots, 5$) refers to the explanatory variables; β_i is the partial regression coefficient; and β_0 is the intercept. $exp(\beta_i)$ is an odds ratio, which mirrors the strength of

the association between the explanatory variables and the remediation probability. When $exp(\beta) > 1$, a positive association exists between the variables and the occurrence probability. When $exp(\beta) < 1$, a negative association exists. When $exp(\beta) = 1$, the variables are not correlated with the event.

Table 5.6 presents the model results. We opt to fit different specifications of the model with a stepwise inclusion of the variables that impact remediation directly or indirectly.

Table 5.6: Coefficients of the logistic regression model for remediation

	Dependent variable: Remediation				
	(1)	(2)	(3)	(4)	(5)
x_1: Mixed notification		1.055^{***}	2.595^{***}	2.876^{***}	3.271^{***}
		(0.292)	(0.644)	(0.649)	(0.758)
x_1: Only email notification		0.695^{***}	0.695^{***}	0.871^{***}	1.192^{**}
		(0.188)	(0.188)	(0.190)	(0.374)
x_1: Only walled garden notification		1.458^{***}	2.821^{***}	3.016^{***}	3.049^{***}
		(0.196)	(0.535)	(0.544)	(0.545)
x_2: Number of walled garden notifications			-1.279^{**}	-1.330^{**}	-1.363^{**}
			(0.457)	(0.466)	(0.466)
x_3: Number of other walled garden notifications				2.320^{***}	2.328^{***}
				(0.280)	(0.280)
x_4: Number of email notifications					-0.234
					(0.249)
x_5: Type of vulnerability					0.289
					(0.222)
Intercept	0.687^{***}	0.412^{***}	0.412^{***}	0.153^{*}	0.130^{*}
	(0.052)	(0.059)	(0.059)	(0.063)	(0.066)
Observations	1,688	1,688	1,688	1,688	1,688
Log Likelihood	-1,076.046	-1,031.975	-1,028.358	-958.041	-956.752
Akaike Inf. Crit.	2,154.092	2,071.950	2,066.715	1,928.082	1,929.504
Note:					*p<0.05; **p<0.01; ***p<0.001

We will first interpret the model following the standard procedure, namely via odds ratios. Next, we will translate the odds ratios into so-called Relative Risks, which probably are easier to understand for readers who are less familiar with odds ratios.

Exponentiating the model's coefficients gives us the odd ratios. Odds ratios express the likelihood of remediation in comparison to a reference group: the control group users (the model's intercept). (Or to be more precise: for models (2) to (4) the reference group (a.k.a. the base category) is the control group, as defined by the categorical variable x_1.

In model (4), we introduce a variable (X3) to control for users in the control group who received other notifications. This does not change the reference group as such, though the intercept shifts down to accommodate the proportional influence on the log-odds of x_3. In model (5), we introduce an extra categorical variable, namely the type of vulnerability. This does change the reference group to control group users with the mDNS vulnerability. This implies that for models (2) to (4) the intercept (β_0) is the mean of the control group defined by x_1, while in model (5) the intercept is the mean of the group that constitutes the reference level for both categorical variables x_1 and x_5: control group users with the mDNS vulnerability.)

The model provides the direction and strength of the association for the predictor variables. Odd ratios above 1 mean that this factor increases the likelihood of remediation compared to the control group, while below 1 implies a decrease. We will interpret the findings based on model (5). It does not perform better than model (4), but it does enable us to look at two additional factors of interest to people designing remediation mechanisms, namely repeated email treatments and whether the type of vulnerability makes a difference. We should note, though, that the vulnerability types are actually technically similar and might show up for the same device. (As it turns out, neither variables have an observable impact on remediation.) Going from model (4) to (5), the coefficients are quite similar. The biggest change is for X1 (email-only). Even in this case, though, the coefficient of model (4) falls within the confidence interval for the coefficient of (5). Based on model (5), we can make the following observations:

- x_1: **Only Email notification**: The coefficient for email-only notifications is 1.19, which can be read as email notifications changing the log odds of remediation by 1.19. After exponentiating the coefficient, this gives us the odd ratio of 3.29 with a 95% confidence interval of [1.57,6.87] (the confidence interval is calculated by exponentiating the confidence interval for the model coefficient). In other words, the odds of remediation increase by 3.29 for users that received only email notifications, compared to the ones that did not receive any.

- x_1: **Only Walled garden notification**: By exponentiating the coefficient value, we obtain an odds ratio of 21.10 (confidence interval: [7.24,62.38]), which indicates an increase of 21.10 in the odds of remediation when notified via walled garden notifications than for not notifying.

- x_1: **Both notifications**: The odds ratio for remediation by users who received both types of notifications is 26.33 (confidence interval: [6.06,120.08]). In other words, using both walled garden and email notifications at least once in different notification days increases the odds of remediation by 26.33.

- x_2: **Number of walled garden notifications**: The coefficient for increasing the number of walled garden notifications for users who did not act upon the first notification is -1.36. This translates into an odds ratio of 0.25 (confidence interval: [0.10,

0.64]), which means we expect to see a decrease in odds of remediation when number of walled garden events increased by one. This is consistent with our findings in Section 6.5, where we observed that remediation rates drop over subsequent quarantining events, indicating that these customers are less able or willing to remediate. Some common reasons why users might not act on the vulnerability notifications are discussed in Section 5.5.

- x_3: **Number of other walled garden notifications**: The odds ratio for remediation for the 192 users in the control group who received notifications for other security issues is 10.25 (confidence interval: [6.15, 18.59]). This means there is a 10.25 increased in odds of remediation compared to those in the control group with no notifications whatsoever. This large positive impact is likely due to the fact that 95% of these other walled garden notifications were made for vulnerable NetBIOS services. The remediation steps are very similar to those for mDNS and OpenResolver. It might even concern the same device that has both vulnerable services running. Disabling the DMZ or removing the device from public access would solve both problems. Thus, we interpret the impact of X3 not so much in terms of a positive learning effect over different notifications, but rather as the effect of sharing the same – or closely related – root cause.
- x_4: **Number of email notifications**: This predictor was not significant. While email-only notifications have a positive influence on remediation, sending subsequent emails did not improve the likelihood of remediation.
- x_5: **Type of vulnerability**: Vulnerability type did not significantly influence the probability of remediation. This might be caused by the fact that both vulnerabilities (mDNS and OpenResolver) require similar actions to fix the problem.

A different way to represent these results, which might be more intuitive to some readers, is to convert the odds ratios into the so-call relative risks (RR). This captures the probability of remediation after the exposure to one of the factors as compared with the probability of remediation in the control group. The RR can be computed as:

$$RR_i = \frac{exp(\beta_i)}{1 - p_0 + (p_0 \times exp(\beta_i))}, \tag{5.3}$$

where p_0 represents the probability of remediation in the control group (i.e., 0.532; see Table 5.2).

Figure 5.5 shows the relative risks computed from coefficients fitted in model (5) using Eq.5.3. Email notifications increase the probability of remediation by 30%, while walled-garden notification push up the remediation probability to 46% as compared to the control group. Adding an email notification on top of the walled garden notification only increases the probability of remediation by a non-significant 1%. (i.e., 47% probability of remediation increase compared to the control group). This suggests that the effectiveness of the mixed treatment is mainly due to the walled garden notification.

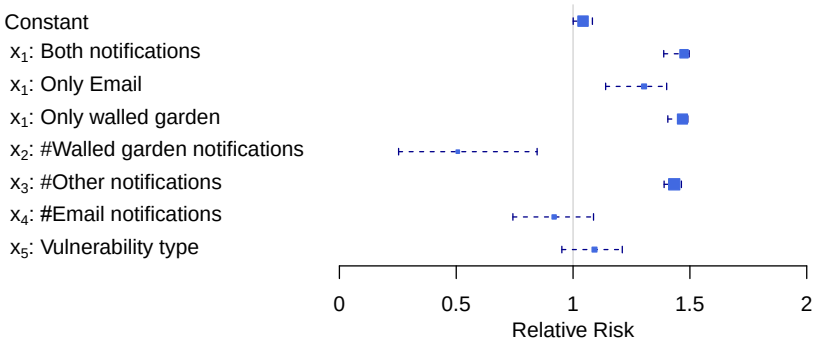


Figure 5.5: Relative risks for each explanatory variable

Subsequent walled garden notifications reduce the probability of remediation by 49%. Users in the control group who received other notifications are 43% more likely to remediate than those who receive no notification whatsoever. The number of email notifications and the type of vulnerability have no significant impact on the remediation probability.

5.5 End user reactions to vulnerability notifications

To gain insight into the user experience of a walled garden or email-only notification, we qualitatively analyzed 324 walled garden forms, as well as 564 emails to the ISP's abuse staff. This corresponds to 384 unique users, 77.4% of the 496 notified users.

We evaluated each message manually with two coders based on a subset of the themes reported by a previous study on ISP notifications[94]. New themes are added where needed. Disagreements between the two coders were adjudicated by a third coder allowing us to solve all conflicts. We found out that issues can be summarized into four categories: (*i*) expressing distrust of the notification; (*ii*) refusing to remediate; (*iii*) asking for additional help or information to solve the problem; and (*iv*) complaining about the disruption caused by the quarantining of the connection. Table 5.7 displays the number of unique users and the percentage of all notified users with at least one message in that category or subcategory.

Table 5.7: Issues raised by users in communication with the ISP

Category	**# unique users**
Distrusts the notification	6 (1 %)
Unwilling or refusing to take action	91 (18 %)
-Does not want to remove settings	80 (16 %)
-Claims to be not vulnerable	11 (2 %)
Requests additional information/help	215 (43 %)
-Requests additional explanation	40 (8 %)
-Requests additional help	169 (34 %)
-Requests a technician	14 (3 %)
-Request to talk with abuse desk	36 (7 %)
-Ask for a retest	38 (8 %)
Complains about disruptiveness	56 (11 %)
-Cannot work due to quarantine	25 (5 %)
-Threatens to terminate the contract	9 (2 %)
-Cannot access devices	34 (7 %)
Other	129 (26 %)
No communication with abuse desk	112 (23 %)

5.5.1 Distrusting the notification

About 1% of the notified users replied to email notifications made by the ISP to check the authenticity of the walled garden or email notifications. These users did not anticipate that their ISP would reach out to them about a vulnerable service. Interestingly, 2 users replied back to the very email they did not trust, to check the credibility of it. The other users contacted the abuse staff to check the authenticity of the quarantine landing page before they followed the suggested steps.

In another prior study, a similar degree of distrust was reported when quarantining broadband ISP subscribers with a malware infected machines[94]. This shows the importance of containing information that allows non-expert customers to reliably tell the quarantine landing page apart from a random phishing page. ISPs might consider personalizing the notification contents to avoid problems such as these.

5.5.2 Unwilling or refusing to take action

A noticeably high number of users did not want to act on the notifications. We distinguish two subgroups of users here. The first group does not want to change the vulnerable configuration of the device. Users argue that the suggested remediation method will prevent them from using their devices and the services that come with it, such as accessing files, playing video games with their friends, or supporting work processes. Thus, they contacted ISP to identify an alternative remediation method. In one specific case, a user mentioned that he paid a technician to set up his modem this way so that he can play games with multiple players. In another case, the user argued that disabling the DMZ and port forwarding will prevent him from monitoring his security camera from outside of his house, rendering his house less well protected. The second group contains users who complain and refuse to take action. They claim that they have been wrongly notified as they took appropriate actions before the notifications. One user claimed that they previously received another notification which also supposedly mis-identified the user's device as vulnerable.

Since a portion of subscribers were unwilling to take action, ISPs might consider withholding the quarantining for subscribers that want to keep their device configurations and suggest that they find an alternative solution to prevent abuse for amplification.

5.5.3 Requesting additional information or help

More than 40% of the users contacted the abuse desk requesting for more information or additional help to solve the problem. This category can be further divided into a few more specific themes: (i) requesting additional explanation; (ii) requesting additional help; (iii) requesting a technician; (iv) requesting to talk with abuse desk and (v) asking for a re-test. Around the first theme, users indicate that they did not properly understand the cause of the problem and requested more information from the abuse desk staff members. Several users indicated that they have been using their devices for years and wanted to know why they haven't been notified previously. Some users misunderstood the security problem and claimed to be secure with a strong login password for intruders. A few users wondered why port forwarding and enabling a DMZ are options on the ISP-issued modems, if these options are now flagged as causing security vulnerabilities. On the second theme, either users could not parse what needed to be done from the notification contents or they had questions about additional remediation methods they could try. About 3% of the users could not solve the problem by themselves and requested a paid technician from the ISP to come and fix the problem. A few indicated that they hoped the technician could find a way to fix the problem so that they can keep their configurations and devices. This rate was much lower than in two previous studies on quarantining broadband ISP subscribers with malware infections.[93, 94]. This might be because patching a vulnerable device is less complicated then cleaning up a malware-infected machine. Around 7% of the users indicated that they prefer to talk to the abuse desk employees over the phone to explain their problem. And lastly, we find out that almost 8% of the users tried to solve the problem

but they were not sure about the effectiveness of the solution and they asked ISP abuse desk members to tell them whether did they managed to remediate the problem.

To reduce the number of requests made for additional help, ISPs can investigate how to improve the usability of notification content. To illustrate: a previous study on IoT malware remediation in a broadband ISP network found that providing more actionable content on the quarantine landing page reduced the percentage of requests made for additional help by half, compared to use of standard content[93].

5.5.4 Complaining about disruptiveness

During the observation period, around 11% of the notified users complained about disruptiveness of the walled garden quarantining. We further investigated the content of these messages and found several recurring themes: (i) customer states s/he cannot work due to the quarantine; (ii) customer states s/he cannot access devices; and (iii) customer threatens to terminate the contract. In the first theme, users indicated that the lack of connectivity means they cannot work from home or conduct their business properly. Around the second theme, users stated that they were out of their homes, or even out of the country, and the quarantining prevents them from accessing their network-attached storage (NAS) systems to access their backups. Finally, around 2% of users expressed anger or frustration and threatened to terminate their ISP subscription. In one case, the user additionally threatened to shame the ISP and their notification procedure on social media. A few users added they were subjected to multiple quarantine events because they could not afford to change the setting or to remove the devices that cause the vulnerability. Some users complained that quarantining users for vulnerabilities are too strong of a measure for this problem.

5.6 Related Work

For many years, a large body of studies has delved into discovering vulnerabilities of different network-level entities namely websites(e.g., [96]), web applications such as CMSes [97], and web infrastructure such as servers [98]. Only in the past ten years have the security research community also put focus on studying the efficacy of notifying affected parties on remediation.

Abuse notifications: Various studies have assessed the impact of abuse notifications on cleanup of compromised websites. Notifications can be sent to the affected owners of the site or to their hosting provider. In an observational study, Li et al. used data of over 700,000 infected websites detected by Google Safe Browsing and found that direct notifications to webmasters via Google Webmaster Console increased the likelihood of cleanup by over 50% and decreased the infection lifetime by at least 62% [33]. Vasek et al. conducted an experimental study on malicious URLs submitted to the StopBadware community feeds to investigate the impact of abuse reports and how the level of detail in the reports

influenced the cleanup rate [32]. They found that abuse notifications sent with detailed compromise information are cleaned up better than those not receiving a notice, 62% compared to 45% after 16 days. Notably, they found that sending a minimal report is roughly as effective as not sending at all. In Chapter 2, we reaffirmed that detailed notices work. We concluded that while around half of all compromised websites were cleaned up after a notification to the hosting provider, sender's reputation played no statistically significant role in the clean up rates [61]. Canali at al. looked into how hosting providers handle abuse notifications [34]. They have notified 22 shared hosting providers regarding their infected webservers and observed that only 36% reacted to the abuse notifications [34]. Similarly, Nappa et al. issued abuse reports for 19 long-lived exploit servers and observed that only 7 providers took action towards cleaning up their malicious servers [35].

Vulnerability notifications: Another branch of studies have looked into how security notifications can expedite vulnerability remediation. For example, Durumeric et al. notified servers receptive to the Heartbleed vulnerability [41]. Through carrying out a controlled notification experiments two weeks after Heartbleed public disclosure, they observed that the patching rates of the notified group was 47% higher than the control group, 39.5% versus 26.8%. Kührer et al. in collaboration with CERTs, clearinghouses, and afflicted vendors notified administrators of vulnerable Network Time Protocol (NTP) servers [42]. Their results indicate 92% of NTP server were patched in 13 weeks time.

Notification mechanism: Several studies investigated specific notification mechanisms. In an earlier chapter, we investigated the usability of walled garden notifications for cleaning malware infections. Our study did not include a comparison with other mechanisms or a control group, which prevented it from measuring the effectiveness of the walled garden compared to less intrusive options [94]. The observed remediation rates were around 70% after the first quarantine event, which is lower than we observed in the current study. The difference might reflect the fact that, on average, infections are harder to remediate than the studied vulnerabilities. As the prior study had no control group, we cannot see to what extent transient events might explain this difference. Such a control group was present in Chapter 6, which studied the cleanup of Mirai infections. The control group did, in fact, show a high rate of transient infection events. Overall, the study found that quarantining and notifying affected customers remediated 92% of the Mirai infections, which is in the same range as the remediation rates found in our study on vulnerabilities. Li et al. studied vulnerability notifications addressed directly to network operators and found them more effective than those send to national CERTs and US-CERT [43]. Stock et al. studied the effectiveness of large-scale email vulnerability notification campaigns. They could only reach around 6% of the affected parties. Of this small fraction, around 40% were remediated once notified [44]. In Chapter 3, we also found email delivery rates to be poor, especially when following RFCs on how to directly contact the resource owner. Stock et al. examined the efficacy of other channels such as postal mail, social media, and phone on remediation rates. Although they resulted in marginally higher remediation rates, the gain from it do not justify the additional costs [45]. Recently, Zhang et al. looked into on the effectiveness

of telephone, email, and instant message (IM) notifications within an ISP with educational institutions as main customers [47]. They conclude that IM is the most appropriate notification mode for such an ISP.

Collectively, these studies investigated the effectiveness of notifications sent to intermediaries as well as the owners of vulnerable servers and websites. However, to the best of our knowledge, there is no prior work that measured the impact of vulnerability notifications sent to end users of residential networks.

5.7 Ethical Considerations

In this study, we leveraged a passively-collected dataset from an ongoing process of vulnerability and abuse handling process by the ISP. All treatments were administered by the ISP. They were existing treatments and took place within the terms of contract with their customers, so no additional consent was needed. We only added the observations from the vulnerability feeds to those treatments. The latter is not regarded as human subject research by our IRB and thus out of scope. Only the ISP's employees could see the customer information that corresponded with each observation of a vulnerable device. The study was conducted on premise at the ISP by one of the authors who was working for the ISP at the time. All raw datasets and the analysis were anonymized. Throughout the study, we followed the policies of the ISP.

5.8 Limitations

We emphasize three limitations associated with our study. First, our findings are tied to the data from a single ISP in Europe. Thus, generalizability and reproducibility of our results to other ISPs or networks are a matter for further research. Second, we only analyzed two vulnerabilities, both tied to devices being used in amplification DDoS attacks. These type of attacks usually are not directed at the vulnerable users themselves. Moreover, there is only limited media coverage of these vulnerabilities compared to, say, Heartbleed or Spectre. These factors may influence the willingness to remediate. Follow-up studies are needed to understand how this impacts remediation rates via quarantining for other vulnerabilities. Third, remediation success is measured from the scan data provided by the Shadowserver Foundation. We assume that these contain the kind of error rates normal for most large-scale scanning efforts. False negatives might lead us to incorrectly identify a host as remediated, e.g., due to temporary network disruptions. We mitigate this issue by only classifying a device as remediated if it did not appear vulnerable in Shadowserver feeds between January 1 - 31, 2018. Last, as we explained in section 5.4.2, there is no way to separate remediation from transient events or DHCP churn. As a result of this, we have overestimated the remediation rates, especially for the control group. This limitation

should not impact our main findings – in fact, this overestimation means the difference with the treatment groups is even larger than we observed.

5.9 Conclusion

We investigated the effectiveness of vulnerability notifications issued by an ISP to its customers in order to remediate devices running open DNS resolvers or mDNS services. After the three month period, we found very high remediation rates for the notified users, especially for the walled garden quarantining and notification: around 87%. These high rates also hold for users who self-released from the quarantine. The email-only notification resulted in remediation in around 75% of the cases. Few studies tracked remediation after three months and in a specific network, so it is difficult to compare these findings to prior work, but the rates are in line with those reported for the NTP amplifier campaign [42].

We explored the relatively high remediation rate for the control group: around 53%, after excluding those customers who received notifications for different vulnerabilities. Several factors cause this rate to be an overestimation. If we would remove all cases where a device was seen only once, we would end up with a remediation rate closer to what other prior studies reported [43, 41]. This would also mean that the difference in remediation rates between the notification mechanisms and the control is likely to be even larger in reality. As it stands, our analysis finds that walled garden notifications increase the probability of remediation by 46% compared to the control group. For email, we find a 30% improvement. However, sending additional walled-garden notifications to subscribers who did not act after the initial notification is associated with a decrease in the probability of remediation by 49%. This indicates certain users are unwilling or unable to remediate the vulnerability.

We have also studied the user experience of these notifications from the communications with the ISP. Quarantining vulnerable device owners is a disruptive treatment. A little over one in ten users complained about the disruption. A fraction of them even threatened to terminate the contract. It is difficult to evaluate this rate of pushback, but it seems a valid conclusion that the ISP is taking the hard road in trying to reduce the security externalities emanating from its network. Other user feedback includes a tiny fraction of users who distrusted the notifications enough to check with the ISP. Almost half of all notified users contacted the abuse department for additional information and help. Less than one in five users seemed unwilling to take action or denied having a vulnerable device to begin with. More actionable notification content might reduce the requests for help and the complaints about disruptiveness[93]. Since writing effective notification content for various vulnerabilities and infections is hard, ISPs could collaborate with researchers to conduct randomized control trials with different forms of content.

All and all, we have demonstrated that quarantining vulnerable devices is a very effective method to remediate vulnerabilities. In the setting of the ISP, email-only notifications

also did much better than in Internet-wide notification experiments and control group. Reachability is likely to be much better, as is trust in the message, given that it comes from the company that users are getting service from.

The high cleanup rates achieved by quarantining and notifying vulnerable resources are comparable to, or even a bit better than, those from prior studies into walled garden notifications for compromised end user devices [94, 93]. This is remarkable, as the vulnerable devices do not pose a threat to their owners, contrary to malware-infected machines.

Notwithstanding these positive results, we do not want to overstate their contribution to solving the challenge of making large-scale vulnerability notifications more effective. The sobering observation that has to accompany our findings is that quarantining is only possible under certain conditions – e.g., the network operator needs to be contractually allowed to do so. More than contractual conditions, though, we expect that many network operators will perceive few incentives to undertake this endeavor. Walled gardens imply direct cost in terms of implementing and maintaining. Then there is the cost of time spent on notifications by the abuse handling staff. Last, but not least, there is the cost of customer pushback.

We should note that the email-only mechanism is cheaper and triggered much less customer pushback and still performed substantially better than the control group. Walled garden notifications achieved an additional 12% remediation compared to the email-only notifications. Is that additional gain worth the higher cost of the walled garden? This is a question for future work. It requires a cost-benefit analysis with the ISP, which is out of scope of the current study.

Still, we do hope that our results will encourage the community to experiment with different mechanisms in order to reach the final goal: realizing the value of large-scale vulnerability discovery for creating more secure networks.

CHAPTER 6

Evaluating effectiveness of ISP-made notifications to users with compromised IoT devices

With the rise of IoT botnets, the remediation of infected devices has become a critical task. As many infected IoT devices reside in broadband networks, this task will fall primarily to consumers and the Internet Service Providers. In the chapter, we evaluated the effectiveness of ISP-made notifications to users with compromised IoT devices. Our main aim is to identify the most effective method to notify the end users against compromised IoT devices. First, we looked into the effectiveness of the walled garden and email notifications compared to natural cleanup. Then, we evaluated the impact of more actionable content compared to standard notification content. We also looked at remediation rates of other networks that did not receive notifications and compare then with our control group to evaluate natural cleanup rates among various networks. Moreover, we investigated reinfection rates. Lastly, we analyze ISP communication logs and conduct interviews with notified end users to understand end users react to the IoT malware notifications made by their ISP.

6.1 Introduction

Events of the past two years have made it abundantly clear that Internet of Things (IoT) devices are being compromised at scale, especially in the consumer space. It is also clear that this situation will not improve in the short term. Due to lack of effective regulations, poorly-secured devices will keep flooding the market. Given the life cycle of the existing and new devices, this means we will be confronted with IoT botnets for years to come.

All this presents us with a critical challenge: how can we remediate the population of vulnerable and compromised IoT devices? Since most of the compromised devices are consumer products, this implies overcoming a number of unsolved problems. A recent study into Mirai [99] identified three critical challenges. First, there is no public information to identify the owner of the device. Second, there is no established communication channel

to reach the owner. Third, where owners are reachable, we do not know how to provide them with an actionable notification. There is often no clear and simple remediation path. In fact, in many cases we cannot even state exactly which of the owner's devices is actually affected.

For the first two problems, identifying and contacting owners, we can turn to an existing arrangement: botnet mitigation by Internet Service Providers. Many of the devices are in access networks, so ISPs can identify and contact the customers who own them. For regular PC-based malware, botnet mitigation by ISPs is widely accepted and has met with some success [77]. However, cleaning up infected devices is still an open problem, even when considering conventional malware. Years of usability research have shown just how hard it is to support end users with little technical expertise in protecting and remediating their personal computers [90]. In the IoT space, all of this becomes much harder. User intuitions ('folk models' [100]) about security are even less aligned with the IoT environment. Furthermore, the actions users need to take are different across devices, vendors and local configurations. Finally, contrary to conventional malware, there are no automated tools to support users in protecting and remediating infected devices. In short, we have no clue whether owners can act at all effectively on the kind of notifications that we can currently provide them with.

We present the first empirical study of the cleanup of compromised IoT in the wild. For this, we collaborate with a mid-sized ISP that notifies Mirai-infected customers via email or by placing their connection in a quarantine network – a so-called 'walled garden'. We measured the remediation rate and speed of 220 users in an observational study and a randomized controlled experiment by tracking the infections in darknet, honeypot and abuse reporting data. We combined this with additional scan data to identify the type of devices that are affected. Next, we studied the user experience by conducting 76 phone interviews and analyzing the logs of the users' communications with the ISP. Finally, we also conducted lab tests with real IoT devices to observe the effectiveness of removal actions and to measure reinfection speed.

In short, we make the following contributions:

- We show that over 87% of all Mirai-infected IoT devices reside in broadband access networks, underlining the critical role of ISPs in IoT botnet mitigation.
- We provide the first real-world measurement of remediation rates for Mirai-infected devices and find that quarantining and notifying affected customers remediates 92% of the infections.
- We find very high natural remediation rates of 58-74% in the control group and in two reference networks where no notifications were sent, probably reflecting the non-persistent nature of the malware.
- We find a remarkably low reinfection rate. Only 5% of the customers who remediated suffered another infection in the five months after our first study. This highlights the effectiveness of the countermeasures taken by the infected customers but stands in contrast to our lab tests, which found very fast reinfections of real IoT devices.

- Remediation succeeds even though customer interviews and communications show that many users are operating from the wrong mental model – e.g., they run anti-virus software on their PC to solve the infection of an IoT device.

Combining insights on the location of compromised IoT devices, effectiveness of different treatments and the experience of real-world users, we contribute scientific evidence for establishing industry best practices around the remediation of compromised IoT.

6.2 ISP botnet mitigation

Cleaning up infected IoT devices can be seen as the next phase of a long-standing challenge: fighting botnets. Over the past decade, mitigation of PC-based malware has consisted of two complementary approaches: taking down the command-and-control infrastructure and cleaning up the infected hosts. Cleanup is an arduous process that demands efforts from different actors, such as operating system vendors, anti-virus vendors, ISPs and the affected end users. As most infected machines reside in consumer broadband networks [77], the role of ISPs has become more salient over time. A range of best practices and codes of conduct have been published by leading industry associations [69, 70], public-private initiatives [71, 72] and governmental entities [79, 74]. These documents share a common set of recommendations for ISPs around educating customers, detecting infections, notifying customers, and remediating infections.

While the existing mitigation practices of ISP are exclusively focused on PC-based malware, they might still provide a good starting point for the remediation of compromised IoT. This assumes, however, that the bulk of the devices reside in the networks of broadband consumer ISPs. To test this assumption, we analyzed the location of compromised devices.

First, following the approach of Antonakakis *et al.* [99], we used darknet data to observe the location of devices infected with a version of Mirai. Darknets, also known as network telescopes, are routed but unused IP address ranges. They passively monitor all arriving traffic at these ranges. We leverage observations from a darknet of approximately $300,000$ IPv4 addresses, spanning 40 networks in 15 countries. As Mirai malware displays worm-like behavior, actively scanning the Internet for spreading itself, we can track its presence in the darknet data. We use data collected in the period January 2016 to April 2018.

We measured per protocol –i.e., per destination port– how many IP addresses were scanning at any point in time. To distinguish Mirai traffic from backscatter traffic and other scanning activity, we uniquely fingerprinted Mirai probes based on an artifact of Mirai's stateless scanning, where every probe has a TCP sequence number – normally a random 32-bit integer – equal to the destination IP address. We observed over 96 million IP addresses. Figure 6.1 shows how they are distributed over six protocols: 23/TCP (Telnet), 2323/TCP (Telnet), 5358/TCP (Telnet), 5555/TCP (TR-069/TR-064), 6789/TCP (Telnet), 7547/TCP (TR-069/TR-064), 23231/TCP (Telnet), 37777/TCP (UPnP), 22/TCP

(SSH), 2222/TCP (SSH), 80/TCP (HTTP), 81/TCP (HTTP), 88/TCP (HTTP), 8000/TCP (HTTP), 8080/TCP (HTTP), and 53869/TCP (Realtek SDK Miniigd). Since Mirai's source code was publicly released, it expanded from targeting telnet to other ports. While port 23 is the second most targeted port, HTTP-related ports have become the main vector – i.e., IoT devices with default credentials for HTTP-related services.

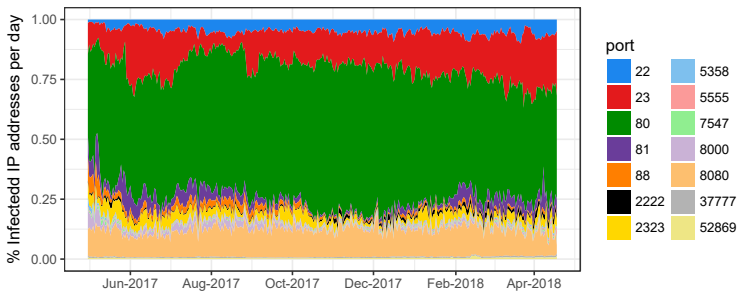


Figure 6.1: Percentage of Mirai-infected IP addresses per port

Next, we mapped these IP addresses to broadband consumer ISPs and other network types. We use the same approach as a study on 'regular' botnet mitigation by broadband ISPs, where a mapping had been developed to identify the Autonomous System Number (ASN) of broadband ISPs in 82 countries [77]. The mapping is organized around ground truth data in the form of a highly accurate commercial database; *TeleGeography Globalcomms* [101], containing market data on the broadband ISPs in 211 countries. In total, $2,050$ ASNs have been labeled manually as belonging to one of the consumer broadband ISPs or to another category: mobile provider, another type of ISP (e.g., business provider), hosting, governmental, educational and other types of networks. Table 6.1 summarizes the percentage of infected IP addresses in each of the network types. The overwhelming majority of these devices (87.61%) are located in ISP broadband networks, while less than 1% reside in other types of networks including hosting, education or governmental networks.

6.3 Partner ISP Remediation Process

Now that we have established that ISPs are in a crucial position to remediate IoT botnets, even more so than for PC-based botnets, the question becomes: what can they realistically do? To answer this question, we have collaborated closely with a medium-sized European ISP with several million customers. The ISP decided to include an abuse feed with Mirai-infected hosts, reported by Shadowserver, in their existing botnet notification and remediation process.

Table 6.1: Distribution of infected hosts across different markets as captured by the darknet (Jan 2016 - April 2018)

#Countries		232	
#ASNs		21 196	
#IP addresses:			
	ISP-broadband	78 885 434	(88 %)
	ISP-mobile	6 888 640	(8 %)
	ISP-other	3 380 164	(4 %)
	Hosting	196 123	(0 %)
	Educational	30 765	(0 %)
	Governmental	313	(0 %)
	Others	655 753	(1 %)
	Total	96 041 559	(100 %)

At the heart of the ISP's process is an industry best practice: placing an infected machine into a quarantine network, a so-called walled garden [78]. There are different ways of implementing walled gardens to fight malware infections. RFC6561 [69] describes two types: *leaky*, an implementation that permits access to all Internet resources, except those that are deemed malicious; and *strict*, an implementation that restricts almost all services, except those on a whitelist. Our partner ISP has implemented a strict version for its consumer broadband subscribers. The walled garden only allows access to 41 white-listed domains, which provide cleanup tools, anti-virus solutions, Microsoft updates, webmail, online banking and a forum for elderly people.

Besides keeping the infected users safely in quarantine, the walled garden also plays an important role in notifying the user. When the user tries to browse the Web, she or he will be redirected to a landing page with a notification about the infection and advice on how to clean it up. The same information is also sent by email to the customers. Whereas emails with the same content can be ignored relatively easily, the walled garden notification cannot.

Next to its own brand, the ISP also provides services to broadband consumers via a subsidiary brand that is targeting the cheaper end of the market. Customers of the subsidiary brand are not quarantined. Notifications are less common and conducted only via email. The ISP also sells subscriptions in the business and mobile service networks. These customers are never quarantined and do not receive IoT related security notifications.

The notification and remediation process starts when an infection is reported in one of the trusted abuse feeds that the ISP receives. For IoT malware, the ISP uses the daily Shadowserver Drone feeds [102]. These include infections labeled as Mirai. The infected machines are discovered through a range of methods, including monitoring sinkhole traffic

and malicious scans to honeypots. If an IP address in the report belongs to one of its consumer broadband subscribers, then the ISP places the connection of that customer in the walled garden. It also sends an accompanying email with the same information. Occasionally, e.g., when the walled garden is full, the ISP sends an email-only notification about the infection.

Once customers are notified via the walled garden, they have three ways of getting out of the quarantine environment. First, they can release themselves by filling out the form on the landing page and report how they have fixed the problem. Submitting the form immediately restores the connection. This option is revoked after two subsequent quarantine events within 30 days, to avoid customers using this route without making an effort to clean up. The second release option is to ask for assistance from the ISP abuse staff to restore the connection. Customers might end up in assisted release because they no longer have the self-release option or because they have contacted the ISP for help. Quarantined customers can contact abuse desk members via email, via the walled garden form, or they can call the regular help desk. The third option is to get a time-out release. After 30 days, customers are automatically released, even if they have not contacted the ISP.

6.4 Study design

Aiming at understanding the impact of the notifications on the remediation process of Mirai-infected devices, we designed a study which consisted of two stages: (i) an observational study on walled garden notifications that the ISP conducted during 4 months; and (ii) a randomized controlled experiment to assess the impact of an improved notification tailored to IoT infection remediation. Figure 6.2 shows the timeline of both studies. Furthermore, to understand Mirai infection dynamics, we also conducted a battery of tests with real vulnerable devices.

6.4.1 Data sources

To identify and track the infected Mirai devices in the ISP network, we leveraged a combination of several data sources. Table 6.2 provides a high-level summary.

Daily Shadowserver abuse feeds

The Shadowserver Foundation is a non-profit security organization that gathers and distributes data on abused Internet resources, most notably compromised machines. It provides network operators with a daily report on compromised hosts in their networks (Botnet-Drone feed [102]). We use the daily reports sent to our partner ISP, in combination with other datasources, to detect and track Mirai-infected users. During the study period, 658

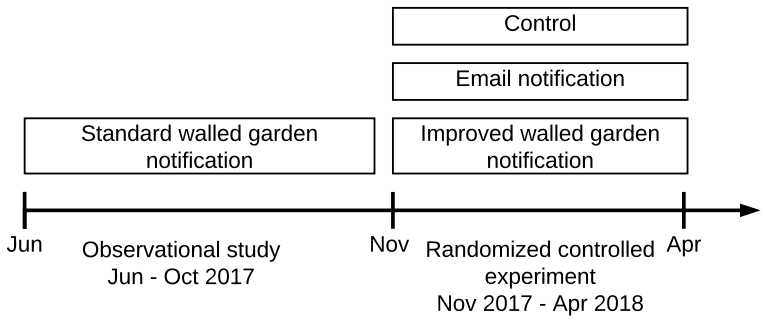


Figure 6.2: Timeline of the experiment

Table 6.2: Data Sources – We used various data sources to analyze the remediation rate of infected ISP subscribers

Role	Data Source	Collection Period	Data Volume
Detecting infections	Shadowserver drone feed	01/06/2017-18/04/2018	658 IP addresses
	IoT honeypot	01/06/2017-18/04/2018	512 IP addresses
Tracking infections	Darknet	01/06/2017-18/04/2018	349 IP addresses
	Shadowserver drone feed	01/06/2017-18/04/2018	349 IP addresses
	IoT honeypot	01/06/2017-18/04/2018	281 IP addresses
Device identification	Censys scans	02/05/2017-16/04/2018	49 Internet-wide scans
	Nmap scans	01/06/2017-18/04/2018	349 port scans
Customer experience	Phone interviews	10/10/2017-18/04/2018	76 subscribers
	Walled garden forms	01/06/2017-18/04/2018	159 forms
	Communication logs	01/06/2017-18/04/2018	521 tickets

IP addresses that belong to one of the ISP's networks were detected as infected with Mirai. We selected 349 of these IP addresses for the purpose of our study (see Section 6.4.3 for the specifics of the selection process). These 349 IP addresses correspond to 343 different subscribers, i.e., there are 6 subscribers whose IP addresses were not completely static during the study period.

IoT Honeypot

An additional data source for detecting and tracking infected devices are the daily log files of a low-interaction honeypot running the open-source IoTPOT software [103]. This IoT-specific honeypot emulates various well-known vulnerable network services by implementing specific IoT architectures. These emulated services include Telnet protocol, IoT devices' HTTP front-ends, the CPE WAN Management Protocol (CWMP) and the remote access setup service of several types of IP cameras. To capture infected IoT devices, the honeypot has been deployed over 738 IP addresses distributed across three countries, including the country in which the partner ISP operates.

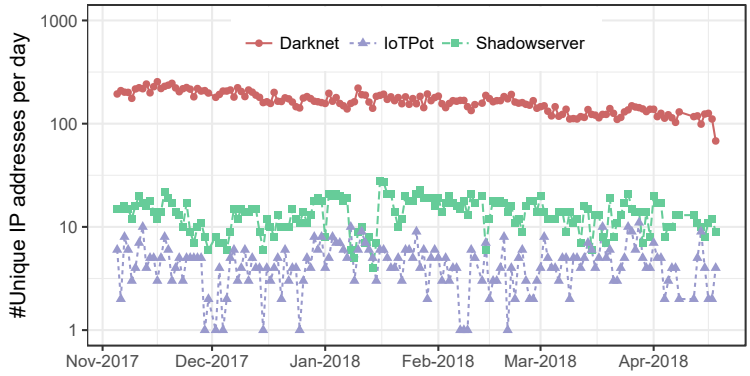


Figure 6.3: Number of unique IP addresses per day of Mirai-infected hosts in the consumer broadband network of the ISP, as detected by Shadowserver, darknet, and honeypot (log-scale)

During the study period, the honeypot captured 512 different IP addresses that belonged to the partner ISP. As the ISP only relies on Shadowserver feeds, we did not use these IP addresses for notification purposes –note that 54.9% (281 IP addresses) of them overlapped with the IP addresses captured by Shadowserver– but instead we used them to track the infections together with the darknet.

Darknet

A third data source for detection and tracking is the darknet mentioned in Section 6.2. We have monitored 16 protocols that are known to be abused by Mirai botnets for the network ranges operated by the partner ISP. The darknet data is much more granular than

the honeypot and Shadowserver data, so we mostly rely on this data for measuring the time to remediation.

Figure 6.3 shows the number of unique IP addresses seen each day in each of the data sources. The darknet has the best coverage, with around 150 unique IP addresses seen every day. The honeypot and the Shadowserver observe only around 10% of these hosts. It is important to note that the ISP's abuse handling process only works with the Shadowserver feed. We use the darknet and IoT honeypot sources only for tracking the infected hosts that entered the ISP abuse handling process.

Censys Scans

Censys [16] is a platform that scans the IPv4 space and aggregates application layer data about hosts on the Internet. We obtained the raw scan data for 49 Internet-wide scans, including HTML code and banner information, for each IP address of the ISP where an infected host was observed. We focused our analysis on scans of CWMP (port 7547), FTP (port 21), HTTP (port 80 and 8080), HTTPS (port 443), SSH (port 22) and Telnet (port 23 and 2323) between May 01, 2017 and April 31, 2018.

NMAP Scans

We used the Nmap network scanner tool [104] to enrich the dataset used for the device identification. Once a device was identified as infected with Mirai, we obtained a list of the open ports as well as banner information. In total, we scanned 349 IP addresses, though 67 of these were already off-line at the time of the scan.

6.4.2 Treatment variables

Our studies are designed to determine the impact of different notification mechanisms on remediation. For this purpose, we compare two experimental treatments using a different notification method (walled garden and email-only) to a control group where no notifications were made during the experiment period. While preparing the experiment, we also improved the standard ISP notification message so as to provide more actionable advice to users. We assess the impact of the improved message via comparing the remediation rate and speed for the new walled garden notification to those measured in the observational study, where the ISP was still using the standard walled garden notification. Figure 6.2 summarizes the different treatment groups that we compare across the two studies. We now take a closer look at the two main treatment variables: notification method and notification content.

Notification method

ISPs have various methods to notify end users for malware infections, such as email, phone calls, SMS, postal mail and a walled garden. However, the efficacy of these methods has rarely been studied, let alone for IoT malware cleanup. In the experimental study, we compare two common methods: email and walled garden.

Email: This method is commonly used by ISPs as it is cheap and easy to scale. However, a major drawback is that it cannot be assured that the email is read in a timely manner, or whether it is read at all. A user might use a different primary email address than the one provided by or to the ISP. The user's email service might also classify the notification as spam. In short, while email is a convenient method, it is unclear how effective this is in terms of promoting IoT malware cleanup.

Walled garden: Walled garden notifications – i.e., the landing page in the quarantine environment – are much more likely to be read by a user. Furthermore, the quarantining provides a strong incentive for the user to remediate. That being said, remediation is not assured. The option of self-release does provide an option to leave the walled garden without any action. Also, when the ISP staff provides an assisted-release, it cannot actually see whether the user successfully remediated. Only when a later Shadowserver report flags the same user again, might the ISP conclude that cleanup failed.

Notification content

Crafting usable security notifications for end users is a difficult challenge. A range of previous studies have focused on how different abuse and vulnerability notification contents can expedite remediation of the security issues [32, 43, 92]. However, such work has not been conducted on remediating IoT malware nor with consumers in real-world broadband networks.

We discussed with the partner ISP the standard notification content that they were using (see Appendix E.1). We noticed it used technical jargon that is probably unfamiliar to most consumers (e.g., Telnet, SSH). Also, the steps that customers were supposed to take were somewhat buried in the overall message. In collaboration with the ISP, we drafted an improved version which avoided certain technical terms and organized the remediation in a numbered series of steps, which we hoped would be more actionable for users. We also added steps to reset the router, as this would close all ports as well as disable the demilitarized zone (DMZ) and universal plug and play (UPnP) (See Appendix E.2).

6.4.3 Study procedure

As shown in Figure 6.2, our study consisted of two stages. The first stage was an observational study of the effectiveness of the existing ISP walled garden mechanism. In the period from June 2017 to the end of October 2017, the ISP quarantined 97 customers and informed them via the standard walled garden notification. All of these users were reported

by Shadowserver as having a Mirai infection. We looked up customer IDs and the set of IP addresses associated with each customer over the period of the study. (Most users retained the same IP address.) We then checked these IP addresses against our three sources of infection data: Shadowserver, honeypot, and darknet. We also logged how long each customer had spent in quarantine, during which they would not be observed in the infection data, of course. By combining these datasets, we could measure remediation success and speed for each user.

Once the first stage of the study was done, we continued with the randomized controlled experiment. To determine the total sample size, in other words how many users needed to be notified, we completed a power calculation for the main outcome variable, remediation rate. We estimated power for an 90% level and used a 10.95 standard deviation based on prior studies [94]. Differences in mean fourteen-day cleanup time of about 10 hours between conditions can be detected with 90% power in two-tailed tests with 95% confidence, based on a sample of 40 Mirai-infected users in each treatment group. This resulted in a total sample size of 120 Mirai-infected users.

The experiment was conducted from the first week of November 2017 to early April 2018. Throughout this period, we followed the procedure summarized in Figure 6.4. First, for each IP address in the Shadowserver report, we identified the customer ID. Then, we checked whether this customer was notified before for Mirai. As prior experience with the notification procedure and remediation actions might influence the remediation time, we discarded a handful of cases that had been notified previously. All others were randomly assigned: 40 users ended up in the walled garden treatment group, 40 in the email-only group, and 43 in the control group. To establish a baseline, the control group was notified later than the treatment groups. For ethical reasons, this delay has to be limited, so as not to expose the customers to unnecessary risks. In collaboration with the ISP, this delay was set at 14 days. After these 14 days, if these customers were still reported in the Shadowserver feeds, then they would be assigned to the walled garden treatment group. When customers in either of the treatment groups were seen again in the Shadowserver feed within this period of 14 days, we would repeat the treatment. In practice this means that some users got multiple e-mails or were quarantined more than once. This study design means the comparison of the treatments will be conducted over these 14 days, though we did keep track of infections and reinfections after this period, as well will explain below.

In parallel to the experiment (November 2017–April 2018), we also collected data on the remediation of infected customers in two additional networks that belong to two different networks of the partner ISP: (i) business services and; (ii) a subsidiary operating under another consumer brand offering broadband services. Customers in these networks do not receive any IoT malware notifications from the ISP. During the experiment period, the business network had 62 infected customers and the subsidiary network had 61 infected customers. We used the same methodology as in the observational study to estimate the remediation rates and compared these to the control group of the consumer network.

Finally, we conducted tests in a lab setup to observe Mirai's infection, cleanup and

reinfection process with real vulnerable IoT devices. By infecting these devices with the malware captured with the honeypot, we could test certain assumptions about removal and reinfection.

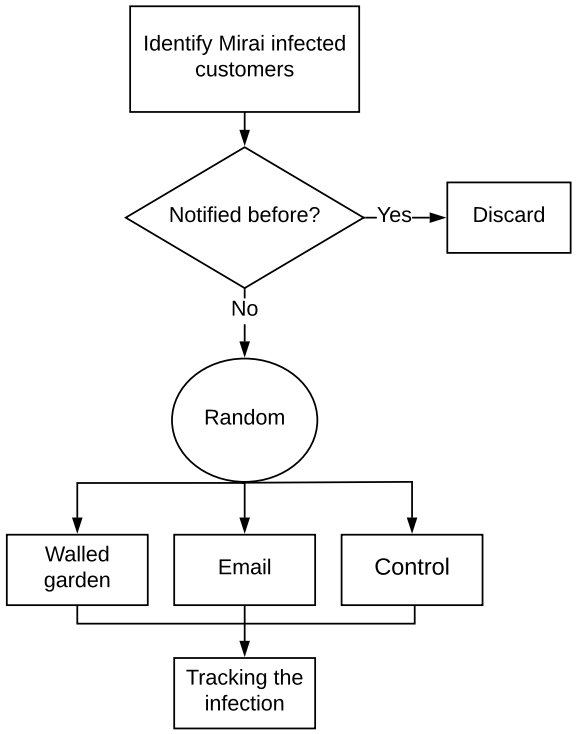


Figure 6.4: Diagram of the randomized controlled experiment

6.4.4 Tracking the infected hosts

Remotely assessing the cleanup status of an IoT device is daunting as passive data sources only allow us to corroborate infections, not cleanup. In this sense, the fact that IP addresses disappear from the infection data (Shadowserver Mirai feed, IoTPOT and darknet) do not necessarily mean the device is clean. We could also be missing observations. It is quite possible for an infected device to not be seen for a few days in the Shadowserver, IoTPOT and darknet data. This can be caused by a range of reasons, including temporary network

disruptions, behavior of the malware or the infected device. (We discuss these limitations in Section 6.9.)

Without additional safeguards, the missing observations during the 14-day period that we track the infections could easily lead us to overestimate the remediation rate. To mitigate this issue, we include a safeguard. After the 14 days, we monitor the infection data sources for an additional 21 days for recurring observations of the customers that were in the experiment. If we see a customer again in this period, we will assume that he has not remediated during the 14 day period. For 34 (15%) of all customers, we collected one or more infection observations in the 21 day period. We therefore set their status to *not remediated* – i.e., still infected – at the end of the 14 days.

Our conservative approach has one downside: within the period of 35 days (14+21), we treat every observation in the Shadowserver, IoTPOT and darknet data as evidence that the infection persists. In reality, some of these cases will be reinfections of devices that had been clean for a short period, rather than continuously infected. In other words, within this period of 35 days we cannot distinguish between infection and reinfection. To reliably measure reinfection rates, we therefore turned to the customers from the observational study. We continuously monitored our data sources for the IP addresses associated with these customers for five months after the observational study period ended in October 2017. If at any point between November 2017 and early April 2018 we saw these customers reappear in the Shadowserver, IoTPOT or darknet data, we would count these cases as reinfections.

6.5 Results

We can now evaluate the effectiveness of the Mirai notifications. As can be seen in Figure 6.5, the total number of Mirai-infected customers was reduced from around 150 to less than 80 infected customers per day at the end of the experiment.

To further understand the impact of the experiment, we will first compare the results for the different treatments (improved walled garden and email-only notification) to both the control group (no notification) and group of the observational study (standard walled garden notification). Next, we will dive into the high remediation rates for the control group. We find similar results in the two reference networks (business and subsidiary brand) where no notifications were issued. We will then discuss the issue of reinfection and long term efficacy of remediation as well as the influence of device type on cleanup. We will end with discussing the results from lab experiments with remediation and reinfection of real IoT devices.

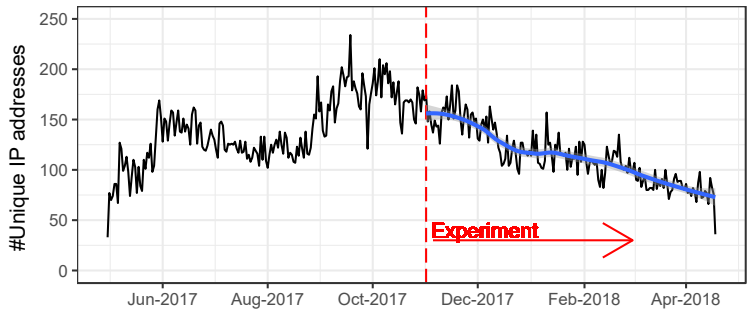


Figure 6.5: Number of infected devices on the ISP's consumer market before and after the notification experiment

6.5.1 Impact of notification mechanism

We first determined the impact of notifications on remediation by comparing the experimental groups. The top of Table 6.3 shows the percentage of IoT devices that were remediated 14 days after the initial notification. It also includes the median infection time for each group. The control group achieved the lowest cleanup rate (74%), closely followed by email-only treatment group (77%). Remarkably, the email-only treatment seemed to have no effect, displaying no statistically significant difference with the control group. The remediation rate of the email-only group is a bit higher, but the median time is a bit longer. The results were significantly better for the customers who received the improved walled garden notifications: 92% of the infected devices were remediated after 14 days. The median infection time is substantially shortened as well: 26 hours, less than half of the 66 hours for the control group.

We also plotted the survival probabilities for the different groups (see Figure 6.6). The groups are quite close one day after the notification, but by day five we see notable differences in the cleanup rates. For instance, 60% of the infected devices in the control group are cleaned within 5 days, compared to 55% of those receiving an email notification and 88% of those receiving improved walled garden notifications.

The log-rank test shows that the difference between the control group and the improved walled garden treatment group is significant ($\chi^2 = 4.4$, $p = 0.0359$). In short, these results provide evidence that quarantining is effective, while email-only notifications are not.

6.5.2 Impact of notification content

To investigate if the improved notification content made a difference, we compared the remediation rates of the walled garden group in the experiment to that in the observational study. Remember, the customers in the observational study were notified with the standard message. Table 6.3 shows a slightly higher cleanup rate and a shorter the median infection time for the improved walled garden treatment group compared to the standard walled garden treatment group. This difference, however, does not pass the log-rank significance test ($\chi^2 = 1.7$, $p = 0.197$). Either the effect is too small to be visible with our sample size or there is no effect. We should also note that this comparison is hampered by the fact that the studies were conducted at different periods in time. In any case, we cannot observe a clear impact of the more actionable walled garden content.

Table 6.3: Summary statistics of Mirai remediation

Groups	Sample Size	% clean	Median infection time	Standard deviation
Control (Experimental study)	43	74%	66 Hours	142.51
Email (Experimental study)	40	77%	74 hours	144.18
Walled garden: improved (Experimental study)	40	92%	26 Hours	91.64
Walled garden: standard (Observational study)	97	88%	27 Hours	121.63
Subsidiary network (Observational study)	61	74%	51 Hours	148.02
Business network (Observational study)	62	58%	198 Hours	141.64

6.5.3 Natural remediation

As we have seen in section 6.5.1, the control group showed remarkably high remediation rates, even though they were not notified.

To confirm the presence of this natural remediation in other networks, we randomly selected 4 other ISPs within the same country where our partner ISP operates and investigated the remediation rates during the period of the observational study. Though we do not control for the potential causes of remediation, figure 6.7 shows that all 5 ISPs actually experience some degree of remediation in their networks even though not all of them issue notifications regarding Mirai-infections. This suggests the pervasive presence of a natural remediation process across different networks. We have investigated potential explanations for this result.

We can rule out three sources of error. First: DHCP churn. Churn often affects measurements that use IP addresses as identifiers for hosts or users. This greatly complicates external tracking, as devices might be assigned new IP addresses during the measurement period. Our results are immune to this problem, as we knew the ISP's customer ID for each

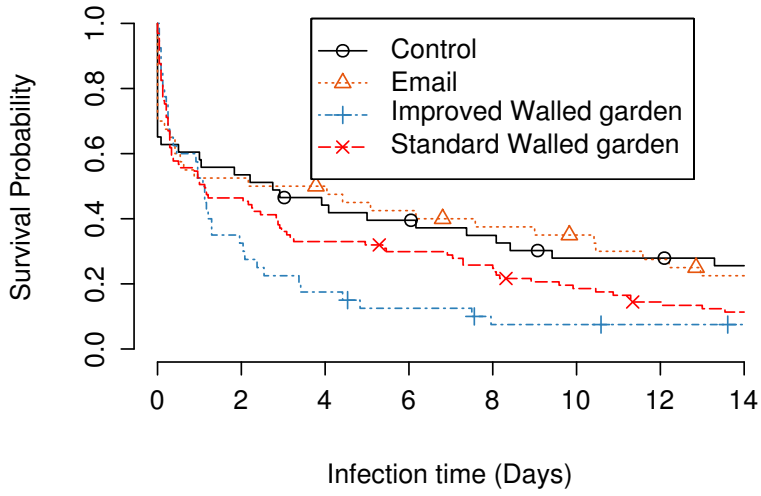


Figure 6.6: Infection rates for the different treatment variables used during the study

user in the study. The ISP's DHCP logs gave us ground truth on the different IP addresses that were assigned to each customer ID over time. Second source of error: additional notifications. If customers in the control group were to receive some other security notification during the experiment, this might trigger remediation actions that could also affect the Mirai infection. Our design, however, ensured that customers in the control group would not receive any other notifications during the 14-day period.

A third source of error we investigated was whether our ability to track infections deteriorated over time. We speculated that perhaps cleaned devices would get reinfected with new Mirai variants or other IoT malware families that we could not observe in the darknet data using Mirai's TCP sequence number artifact. While theoretically we cannot rule this out, we do observe that overall Mirai infection levels remained more or less constant in the darknet data. So the Mirai variants that produced the initial infections were still very active. There was even an increase in command-and-control servers reported during that period [105]. Also, we saw none of the affected customers reappear in the other two datasets: Shadowserver and IoTPOT.

One explanation that can explain, at least partially, natural remediation is the fact that Mirai infections are reported to be non-persistent [106]. We also confirmed this ourselves (see section 6.5.7). This means that every power cycle or unplugging action leads to cleanup. High natural cleanup might thus be driven by users who turn off devices or otherwise disconnect them, rather than use them continually. Indeed, many of these infections

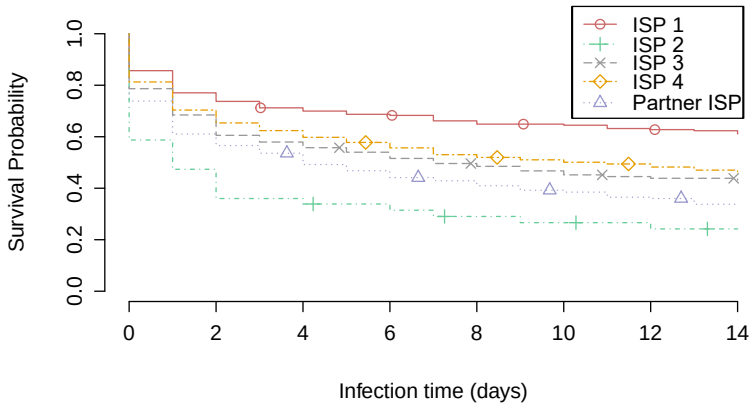


Figure 6.7: Cleanup rates for 4 randomly chosen ISPs within the country where the partner ISP operates

are very short-lived. Around 37% of the infections in the control group are seen only once or twice and disappear from the darknet data within one hour. These transient infections might also reflect volatile usage patterns specific to certain IoT devices. Think of a NAS device that is temporarily connected to another network, perhaps at a friends house. It gets infected there, but then is removed again from the network.

Now, these devices might get cleaned naturally because of usage patterns, but wouldn't they quickly get reinfected again when they are turned back on? In the experiment, we cannot distinguish between infection and reinfection (see Section 6.4.4), so this might happen. However, all the devices that we counted as clean were not seen again for 21 days after the experimental treatment ended. This suggests that reinfection stopped at some point. Something must have changed, beyond a mere reboot. We take a closer look at the issue of reinfection in section 6.5.5.

6.5.4 Natural remediation in other networks

To investigate whether the high natural remediation rate in the control group was an idiosyncratic result specific to this network or customer base, we also analyzed the infection data for two other networks of the same ISP: their business services network and the network of a subsidiary brand offering consumer broadband on the cheaper end of the market. We compared the remediation rate of the control group from the experiment to the rates for the two other networks. As with our control group, the customers in the two other networks did not receive any notifications for IoT infections from the ISP. This makes them

very relevant points of reference.

As shown in Table 6.3, the other networks also display high natural remediation rates. The rate in the business network (55%) was lower compared to the control group (74%) and the subsidiary (74%). Remediation in the two consumer groups (control and subsidiary), however, are virtually the same. Figure 6.8 also shows this pattern. The log-rank test reports a significant difference between customers with business service subscription and the control group (log-rank test, $\chi^2 = 5.4$ with $p - value = 0.0196$) and business network and subsidiary network (log-rank test, $\chi^2 = 4.9$ with $p - value = 0.0268$).

The median infection time for the business network was also significantly longer compared to the other networks. One hypothesis for this finding is that for business continuity reasons, business customers are less likely to reboot or power off their devices as often as consumers. Related to this different usage pattern, we would also expect the composition of IoT device types to be different from the two consumer groups. As we will discuss in section 6.5.6, this is in fact the case. Taking these factors into account, we find very consistent natural remediation rates across the different networks, increasing our confidence in the results of the experiment.

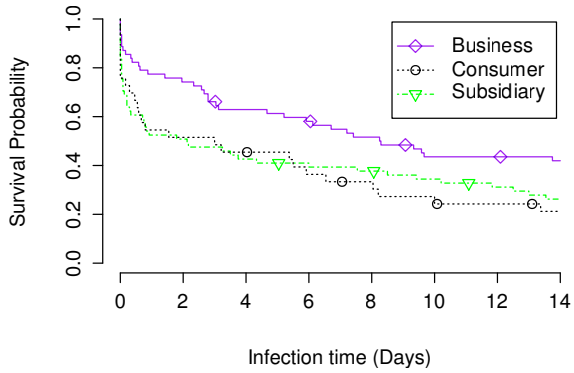


Figure 6.8: Survival curves of the Mirai infections

6.5.5 Long-term efficacy

The non-persistent nature of Mirai means that rebooting, shutting down or unplugging an infected device would cause it to be removed. This fact seems to be an important driver of the high natural remediation rate we observed during the experiment. However, merely rebooting the device does not fix the underlying problem as the device remains vulnerable to infections once it comes back online. To put it differently, the high remediation rates we

observed in our experimental and observational study might be Pyrrhic victories if the devices are simply reinfected again soon thereafter. Removing the underlying problem would require affected users to take other actions, such as changing default passwords, updating the firmware or changing router settings – measures that are much more complicated than a mere reboot.

To get a sense of reinfection rates and the long-term efficacy of remediation efforts, we looked at the 97 customers in the observational study. We investigated reinfection rates for this group during a five-month period after the initial 35 days tracking period. We find that only 5 of these customers (5%) were seen again at some point during those five months in the Shadowserver, IoTPOT, or darknet data. In other words: not only is short-term remediation very high, the longer-term reinfection rate also is surprisingly low. This strongly suggests that whatever action the customer took, it was more than a mere reboot of the device. We have asked users about the actions they took and discuss the results in Section 6.6

On the other hand, intentional action by the user cannot explain the whole story. This is what the high natural remediation rate in the control group tells us. The high remediation rate also contains a signal about low reinfection rates. Remember that to conservatively count them as clean, we tracked the customer IP addresses for an additional 21 days. We did not see these devices again, which clearly means they stopped getting reinfected at some point. In other words, while we might explain the quick removal of Mirai from the combination of non-persistence and device usage patterns, this does not explain why most devices are never seen again. In short, while the low reinfection rate is a positive finding, it is also one for which we have no explanation.

6.5.6 Impact of device type

So far we have encountered a number of surprisingly positive results: high remediation rates across all groups, even in the control group, the two reference groups, and low reinfection rates in the months thereafter. To understand if these results are somehow the result of a peculiar composition of device types in these networks, we take a closer look at the affected devices. Is there anything special about them in terms of the cleanup actions or usage patterns?

Following a similar methodology as Antonakakis *et al.* in [99], we have used Censys [16] to determine the device types. We analyzed the banner information obtained through Censys scans and were able to label 88 devices (28%). These devices were mainly network cameras/DVRs (11%), storage units (7.44%) and routers (3.83%). However, the Censys scans did not allow us to label 72% of the infected devices due to the lack of banner information. In order to increase the number of identified devices, we further conducted port scans on the unidentified devices suing the Network Mapper (Nmap). With this active scanning we gathered banner information of additional ports, i.e, port 5000 (UPnP), 8443 (alternative HTTPS), 32400 (Plex media) and 37777 (QSee DVRs). This allowed us to label

36 additional devices.

Table 6.4: Type of infected devices per service

Service	Device type	Amount of Devices	
FTP	NAS	20	(8 %)
	Router	13	(5 %)
	Server	3	(1 %)
	Set top box	1	(0 %)
Telnet	Set top box	6	(3 %)
	DVR	4	(2 %)
HTTP	Camera	13	(5 %)
	DVR	5	(2 %)
	Printer	4	(2 %)
	NAS	3	(1 %)
	Media streamer	2	(1 %)
	Server	1	(0 %)
HTTPS	Media streamer	3	(1 %)
UPnP	NAS	9	(4 %)
Alt. HTTP	Camera	18	(8 %)
	Media streamer	1	(0 %)
	Firewall	1	(0 %)
Alt. HTTPS	Router	11	(5 %)
Plex	Media streamer	1	(0 %)
QSee DVR	DVR	3	(1 %)
	Total identified	124	(36 %)
	Unknown	219	(64 %)

Table 6.4 shows the types of devices identified by port. The devices we identified were primarily network-attached storage (NAS) appliances, home routers, cameras, DVRs, printers, and media streamers. This composition of device types is consistent with the composition reported in an earlier study on global Mirai infection [99], suggesting our findings are not driven by selection bias in the types of devices that were affected and remediated.

Device type does seem to influence the infection time. Figure 6.9 shows the survival curves for the top 5 most common types of devices in our study. The results show that around 50% of the DVRs and cameras remain still infected, while only 20% of the infected routers and NAS appliances were infected after 14 days. While these overall remediation rates per device type seem to indicate that some devices are easier to clean, the survival

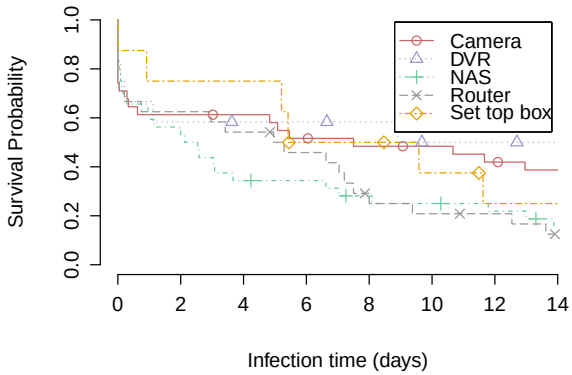


Figure 6.9: Cleanup rates for the top 5 device types

curves did not show significant differences (log-rank test, $\chi^2 = 7.1$, $p - value = 0.1$).

Interestingly, the composition of device types was different for the business network compared to the two consumer networks (see Figure 6.10). Routers, security cameras and videoconferencing hardware were more common in the business networks, while storage units and DVR were mainly present in the customer and subsidiary networks. This supports our earlier speculation that the natural remediation rate is indeed tied to the usage patterns of the devices. Remember that the natural remediation rate in the business network was lower. We now see that indeed this concerns a different device population. More of these devices are likely to be always-on for business continuity reasons. If rebooting or unplugging occurs less frequently, there is also less opportunity for natural remediation to occur.

6.5.7 Lab testing of cleanup and reinfection

In addition to the observational study and the randomized controlled experiment, we also conducted a series of in-lab tests with actual vulnerable devices. These simple tests aim to test the assumption that Mirai malware was indeed not persistent and to also shed some light on reinfection.

The test environment consisted of 7 vulnerable devices (1 IP camera, 1 printer, 1 home router, 3 network storage units, and 1 satellite TV receiver) in their default state (i.e., with their network ports open, and able to accept default credentials). We infected them with a Mirai binary captured by the honeypot. Once infected, we connected the devices to the public Internet and logged all the incoming/outgoing traffic. After malicious outgoing traffic was observed in the infected devices, we rebooted them. Our results showed that after the restart there were no signs of infection in any of these devices: (i) no suspicious

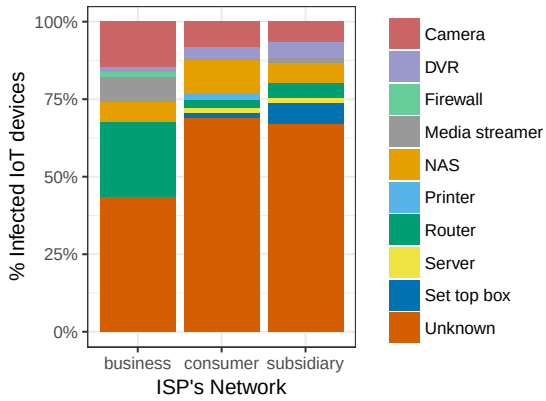


Figure 6.10: Distribution of device types per network

process was running after the reboot; and (ii) no malicious communication traffic was observed. However, even though the binary was not running in any of the devices, we did find it in the file system of one of the devices as this device was using a non-volatile storage and the presence of the malware file survived the reboot.

These results are in line with previous studies [106] which also demonstrated the non-persistent nature of Mirai infections. (While [99] did report some persistence, this appears to be related to binaries for X86-64, so non-IoT.) In general, our findings confirm the advice to consumers to reboot the device, though this alone does not resolve the underlying vulnerability. As long as non-persistence is the norm, rebooting will remain effective. As recent as May 2018, the FBI issued a global alert with the same advice [107] for dealing with a massive population of devices compromised with VPNFilter. Of course all of this, including the high remediation rates we reported earlier in this section, will change when attackers find a way to gain a more persistent foothold on the devices. There are early signs that this is happening [108].

Next, we investigated the reinfection rate, i.e., the time it takes to infect a device, that was cleaned, again. To this end, we connected the devices back to the Internet after rebooting them and monitored the outbound traffic to see whether they get reinfected. We conducted the same procedure three times for each device. Table 6.5 shows the average reinfection speed per device. Five out of six devices got reinfected within an hour after being rebooted. This high reinfection rate is consistent with the aggressive scanning behavior of Mirai. One vulnerable device did not get reinfected. A closer analysis of the traffic showed that indeed there were infection attempts but the implementation of the telnet service denied any login attempt for 30 minutes after an unsuccessful login attempt. The

aggressive scanning behavior together with the timeout of the telnet service served as an impediment to reinfection.

Table 6.5: Reinfection rate per device type

Device type	Mean time to reinfection
IP camera	No infection for 48 hours
Printer	19min 0sec
Router	1min 50sec
NAS 1	14min 35sec
NAS 2	47min 9sec
NAS 3	37min 47sec
Satellite TV Receiver	5min 35sec

These results have two implications for our study. First, it underlines the validity of the conservative approach that we took in measuring remediation. Our tracking methodology did not allow us to measure reinfections on a granularity of minutes. This means it is not feasible to distinguish infection from reinfection. It makes more sense to collate the different infection observations over time into a more or less persistent status of being infected.

Second, and more important, this aggressive reinfection behavior means that if we do not see a device for 21 or more consecutive days (our extended tracking period, see Section 6.4.4), then some remediation action was taken that goes beyond a mere reboot. No vulnerable device with a direct connection to the Internet would survive that long without reinfection.

6.6 User experiences

Our experimental results show remarkably high remediation rates, especially for the improved walled garden notification. While this is a hopeful result, it is also truly puzzling. We know from prior work that remediation is difficult for end users, even for the more conventional scenario of cleaning up PC-based malware (see related work, Section 6.7). In this scenario, it is easier for the user to identify the offending device and the ISP can tell the user more precisely what steps she or he needs to take and point to readily available tools to automatically detect and remove the infection. In other words, the notification is much more actionable for the user.

Compared to the conventional scenario, remediating IoT malware seems much more difficult for users. Even in our improved notification we cannot tell the user which of their devices is affected or even what type of device they should look for and disinfect. Next,

there are no tools available for disinfection. Finally, remediation actions vary greatly per device type, vendor, local configuration, etc. Absent all of this information, the notification is limited to describing several rather generic actions. And yet, we find very high clean-up rates – higher, in fact, than the rate for PC infections. We have a direct point of comparison from a prior study conducted recently also at a European mid-sized ISP [94].

The high remediation rate puts a premium on better understanding how users responded to the notification. In this section, we analyze data on the user experience of IoT cleanup collected via phone interviews and the communication logs of the ISP.

6.6.1 Phone interviews

We called 173 customers to invite them to participate in a short telephone interview. This includes all customers in the observational study and the experimental study, except for the customers in the control group and 4 customers who had terminated their contract in the time between the treatment and the interview.

In total, 76 (44%) of the customers accepted the invitation. The response rate was nearly the same in each treatment group. The non-response consisted of customers who did not want to participate (20, 12%), or who could not be reached by phone within several attempts (77, 44%).

Table 6.6: Respondents receiving and reading the notification

Experimental group	Total	Received	Read	Distrust
Email-only	16	8 (50 %)	6 (38 %)	2 (13 %)
Walled garden (improved)	18	18 (100 %)	18 (100 %)	0 (0 %)
Walled garden (standard)	42	40 (95 %)	36 (86 %)	6 (15 %)

We first asked participants if they remembered receiving the notification and, if so, if they remembered reading it. Nearly all customers in the walled garden groups remembered receiving it, compared to just around half of the customers in the email-only group. For those customers who did not remember receiving the notifications, we checked whether we used the correct email address. All confirmed it was correct. In other words, the emails likely reached their inbox, but were overlooked (or perhaps got caught in the spam filter). Most of the customers who remembered receiving the message also remembered reading it (See Table 6.6). Some of the customers who did not read it mentioned that they did not trust the message and wondered whether it was a phishing mail. (One interviewee also did not trust our phone interview and thought it was a Microsoft scam call).

We then asked the 60 customers who remembered reading the notification if he or she took any action and, if so, what action. Four respondents (6.7%) said they did nothing. A further seven (11.7%) said they had called an IT repair service and did not know what

this person had done exactly. All others listed doing one or more of the steps mentioned in the notification, most often mentioning their attempts to identify the offending device. Furthermore, 22 customers (36.7%) specifically stated they had disconnected a device like a camera, DVR or NAS device from the network. One even claimed to have thrown the device in the trash. Also, 22 (36.7%) people mentioned changing the password for one or more devices and 23 (38.3%) said they reset one or more devices. One customer mentioned conducting a firmware update. Four customers reported that they had run an anti-virus scanner. This latter answer signals a misunderstanding of the nature of the infection. We encountered this more frequently in the communication logs, which we discuss below.

Next, we asked whether the customer sought additional help for the problem. Thirteen people (21.7%) mentioned seeking help from another person, such as their relatives or calling the ISP's help desk. Ten people (17%) asked the ISP to send a paid repair person and one person contacted another repair service. Another form of additional help is searching the web. Five people (8.3%) used Google to find additional information and one person mentioned that they consulted the website of the manufacturer of the offending device.

76 respondents were asked how confident they felt in their ability to solve computer security issues like this. Surprisingly, the largest number of people reported to be very confident (34%) or fairly confident (29%). Some of these respondents elaborated on their answer by stating that they had competent people in their environment who they could turn to. On the other end of the spectrum, 17% ranked themselves as not very confident and 18% stated having no confidence at all and little to no knowledge about these issues. Several of these people said they always ask someone else for help. Some of the interviewees stated that they considered themselves too old for these types of problems. We analyzed the correlation between confidence level and cleanup success and found no relationship. It seems confidence, or lack thereof, does not predict remediation outcomes.

We ended the interview by asking all customers how the ISP can improve its communication about these issues with customers. This question revealed wildly different experiences. On the positive side, 17 respondents (22%) explicitly stated being satisfied or even very satisfied with how the ISP handled the situation. A few suggested sending prior warnings before quarantining the connection and to provide more specific information on what to do and what the offending device is. Another suggestion was to provide an option to contact abuse staff during evenings or weekends for customers who cannot self-release from the walled garden. On the negative side, nine respondents (12%) expressed dissatisfaction or anger about the process. The most vocal critics said that they had incurred economic losses as they were running small businesses on their consumer subscription which were interrupted by the quarantine event.

6.6.2 Communication logs

Additional insights into the user experience of IoT cleanup were extracted from the communication logs between the help desk and the customers in the study, except for those in

the control group. In total, we found one or more messages for 92% of these customers in the ISP's logs. We investigated 159 walled garden contact forms (from 90 unique customers), 404 emails (from 106 unique customers) and 117 help desk logs (from 68 unique customers).

First, we explored the distribution of messages across the different treatment groups (See Table 6.7). We found that about a third of the customers replied to the email notification and only 3 customers contacted the help desk. This rate is much higher for the walled garden groups: around 50% of the quarantined customers called the help desk. While less communication is cheaper and improves the incentives of ISPs for cleanup, it seems that the rate of seeking help is related to action on the side of the customers. As we saw in Section 6.5.1, the remediation rate of the email-only group was indistinguishable from the control group. The walled garden groups did take action and this is also associated with more communication with the ISP.

Table 6.7: Communication channel used by customers in different groups

Experimental group	n	email	contact form	helpdesk
Email-only	40	16 (40.0%)	–	3 (7.5%)
Walled garden (improved)	40	23 (57.5%)	31 (77.5%)	21 (52.5%)
Walled garden (standard)	97	67 (69.1%)	59 (60.8%)	44 (45.4%)

Next, we read a sample of about 20% of messages in each category and created labels for recurring themes. We then read all messages and manually labeled each one as to whether a certain theme was present in it or not. Table 6.8 presents the results aggregated over all customers, i.e., whether a theme was present in one of the messages of a customer. The general pattern confirms what we found during the phone interviews. Some issues are more salient, though. In the walled garden treatments, about one in three customers states that they have run an anti-virus scanner on their PC to remediate the problem. This underlines, even more than the phone interviews, that a significant portion of affected population does not understand the basic properties of IoT malware, even when they have actually seen and read the notification. We found a weak correlation with remediation: customers who mention running anti-virus remediated more slowly. Around 60% was clean after five days, whereas 60% of the other customers was clean within little more than one day. That being said, both groups reached 90% remediation in two weeks.

While a significant portion of the users is working from an incorrect mental model ('folk theory' [100]) of the problem, they do seem to be able to remediate in the end. Of the 51 customers that mentioned running a virus scanner, 23 also mentioned disconnecting a device. Proportionally, this rate is actually a bit higher than for the people who did not mention running anti-virus. Overall, around 40% of the customers in the walled garden groups mention that they disconnected a device, compared to just 7.5% for those who

received the email-only notification.

In the improved walled garden group, dissatisfaction or frustration is substantially lower than in the standard walled garden group. We are not sure how to explain this. It might be that the improved message is more helpful. We should note, however, that the improved walled garden notifications were issued several months later in time than the standard notifications. By that time, more people might have seen reports in the media about IoT compromise and they might thus be more accepting of the need to take countermeasures.

Table 6.8: Themes of user experience in communication with the ISP

	Email-only	Walled garden (improved)	Walled garden (standard)
	n=40	n=40	n=97
Runs a virus scanner	7 (17.5%)	12 (30.0%)	32 (33.0%)
Identifies IoT device	9 (22.5%)	17 (42.5%)	58 (59.8%)
Requests additional help	2 (5.0%)	8 (20.0%)	41 (42.3%)
Wants possibility to call the abuse team	0 (0.0%)	2 (5.0%)	16 (16.5%)
Requests paid technician	0 (0.0%)	4 (10.0%)	11 (11.3%)
Disconnects device	3 (7.5%)	15 (37.5%)	42 (43.3%)
Cannot work due to quarantine	0 (0%)	4 (10.0%)	18 (18.6%)
Complaints over disruption of service	0 (0%)	1 (2.5%)	13 (13.4%)
Threatens to terminate contract	0 (0%)	1 (2.5%)	5 (5.2%)

All in all, the customer experience data helps us to make sense of the high remediation rates for the walled garden groups. While users might not grasp the technical foundations of the infection, as signaled by running AV on a PC in their network, they still end up taking effective action. Disconnecting devices is an intuitive countermeasure, after all, even if it is also costly on the side of the customer – in the sense of not being able to use the device.

It is tempting to speculate about how these customer responses might help explain the remarkably low reinfection rate of the the standard walled garden group (see Section 6.5.5). One might reason, for example, that these users either keep the devices disconnected over a longer period or that they reconnect them differently than before, leaving them no longer exposed to the public Internet. Another explanation is that they factory reset their router, which for certain models means closing open ports and disabling the DMZ and uPnP. This leaves the user in a less vulnerable state.

In the end, though, these speculations seem somewhat beside the point. Remember,

even the control group had a low reinfection rate, in the sense that most of the customers in that group were not seen again for at least 21 days after their initial infection disappeared (see Section 6.5). Whatever the explanation is for this result, it could very well also explain the bulk of the low reinfection rates in the other groups in the study, rather than intentional remediation actions on the side of the users in those groups. For now, we are stuck with a mystery that future work will have to resolve.

6.7 Related Work

In this section, we briefly review three related areas of work. We survey studies on botnet mitigation by ISPs, efficacy of abuse notifications and end user security behavior.

Botnet mitigation by ISPs: Various studies have looked into the role of ISPs in the fight against botnets mitigation and remediation. Most notably, [77] empirically confirmed the point that ISPs are indeed critical control points for botnet mitigation and that infection levels are very different across ISPs, even when they operate in the same country and market, demonstrating they have leeway to act. Work on Conficker cleanup [76] found no clear impact of national initiatives to mitigate botnets.

Additionally, industry groups and international organizations have published ISP best practices that explain how to contact and clean up infected customer's machines. RFC 6561 describes various methods that can be used by ISPs to notify end users about a security problem [69]. Some of the described methods include postal mail, email, phone or walled garden notifications. On the other hand, the effectiveness of the methods is not discussed in detail. A report outlined by M3AAWG identifies best practices for walled garden notification. However, ITU's Anti-Botnet Toolkit raised potential issues that may result from ISP notifications [79].

In an earlier chapter, we investigated the usability of walled garden notifications. This study mainly focused on regular malware infections of PCs while presenting a simple comparison on remediation rates per malware type which also includes Mirai. This study was purely observational and we mainly analyzed users' behavior while in quarantine. Nevertheless, we reported overall remediation rates as observed for the whole system which cannot be solely attributed to the effect of the walled garden. We found that were roughly over 85% of Mirai-infected machines were cleaned after 2 weeks period, which is a bit lower than both standard and improved walled garden notifications for Mirai-infected customers observed in the current study. On the contrary, in Chapter 6, we focused on analyzing the actual impact of the walled garden by designing an experiment with a control group which allowed us to estimate the efficacy of the walled garden notifications on their own. Moreover, this study is specific to Mirai-infected devices which allowed us to customize the content of the notifications with IoT-specific cleanup instructions.

Efficacy of abuse and vulnerability notifications: A large body of research focuses on efficacy of email notifications on large scale vulnerability notifications. For instance,

Li *et al.* issued various types of notifications to CERTs and operators of networks [43]. They concluded that detailed notifications to operators made the highest impact. On the other hand, their results suggested that overall vulnerability remediation was marginal, even with detailed notifications to operators. Similar to this work, Stock *et al.* studied the feasibility of large-scale vulnerability notification and found that notified parties achieved higher remediation rates than the ones that received no notifications [44]. Additionally, in Chapter 3, we demonstrated the poor deliverability of email-based notifications and proposed searching for other mechanisms to deliver notifications. Stock *et al.* evaluated the effectiveness of other mechanisms to delivery vulnerability information such as postal mail, social media, and phone and reported slightly higher remediation rates for these mechanisms [45]. On the other hand, they stated that slightly higher remediation do not justify their costs and additional work put into issuing them. Majority of these studies used email to reach affected parties. Because it scales reasonably well. Conversely, many emails bounced before even reaching the affected parties. Moreover, the ones that reached often triggered no follow-up actions.

Another series of studies explored the efficacy of email notifications on abuse remediation. These notifications are sent to the affected owners of the site or to their hosting provider. Li *et al.* assess the influence of abuse notifications for $761,935$ infected websites detected by Google Safe Browsing and Search Quality [33]. Direct notifications to webmasters increased the likelihood of cleanup by over 50%. Vasek *et al.* found that verbose notifications to webmasters and hosting providers were the most effective [32]. In Chapter 2, we studied the effect of reputation of the sender of the abuse notification on cleanup rates. While notifications in general improved cleanup, there was no observable effect of the sender reputation.

Results of these website cleanup studies indicate a much lower remediation rate than that observed in this study. This could be partly because of Mirai's non-persistent nature.

End user security behavior: A large body of work has studied the challenges of end users in obtaining and following security advice. A study on end user perceptions on automated software updates concluded that most users do not correctly understand the automatic update settings on their computer and thus cannot manage to update as they intend to [88]. Fagan *et al.* [89] investigated user motivations regarding their decisions on following common security advice. They reported that the majority of users follow the usability/security trade-off. Forget *et al.* collected data on users' behavior and their machine configurations and highlighted the importance of content, presentation, and functionality of security notifications provided to users who have different expertise, expectations, and computer security engagement [90]. This work demonstrated the importance of effective communication between customers and the ISP. This can help to ensure a better understanding of the notifications and a higher rate of remediation.

6.8 Ethical Considerations

This study leveraged passively collected datasets and a small number of active scans that were carried out following the guidelines of the Menlo report [109]. All raw data and statistics generated during the study were anonymized, and only the partner ISP's employees knew what customer corresponded to which infection. We always followed the policies of the ISP and notified all the infected subscribers accordingly. We only added the experimental design of random assignment and the observation in abuse feed and darknet data of the infected devices. The latter is not regarded as human subject research by our IRB and thus out of scope. For the purpose of the experiment, the customers in the control group received the notification with a delay of 14 days. Moreover, during the phone interviews, interviewees were provided with an opt-out option. Throughout the interview process, only 20 interviewees asked to be excluded from the interviews.

6.9 Limitations

Our study faces three key limitations. First, detecting and tracking infections is difficult. No method detects all infected machines and when tracking a detected infection there will also be missing observations complicating inferences about cleanup. The former issue is less of a problem for our study, as our design is not based on capturing all infections. The latter issue we mitigated by adopting a very conservative approach in measuring cleanup. If we saw the same customer again within 21 days after the experiment, we would assume they were not cleaned up, irrespective of the missing observations in between. This gives us a lower-bound estimate of cleanup.

Second, the external validity of this research project is open to discussion. On the one hand, the study is conducted in a real-world setting within normal business processes. In addition, the ISP is the second largest in the country and has several million broadband customers. They represent a wide variety of people in terms of demographics. Therefore, we have no reason to assume that our findings are particular to this ISP. On the other hand, it is impossible to know to what extent a walled garden mechanism at another ISP would get the same results until follow-up experiments are conducted.

Last, the dynamic nature of malware limits the generalizability of our findings. Our results are based on Mirai. During the study period, new Mirai versions and other IoT malware families were still non-persistent. This greatly increases natural cleanup via rebooting of devices and it also facilitates cleanup by end users. As IoT malware becomes more sophisticated, it seems a matter of time before they are able to establish a more permanent foothold on the device. Indeed, a recent study reported the first persistent IoT malware [108]. We expect this to cause lower remediation rates.

6.10 Conclusion

We have presented the first empirical study on the cleanup of IoT malware in the wild. We found that quarantining and notifying infected customers via a walled garden remediates 92% of the infections within 14 days. Email-only notifications have no observable impact. We also found high natural remediation rates and low reinfection rates. We have no good explanation for the low reinfection rate, though we are quite confident the result itself is correct. While quarantining infected devices is clearly highly effective, future work will have to resolve these remaining mysteries.

At first glance, the implications of our study for industry seem clear. First, ISPs have a critical role to play as more than 87% of the infections reside in their networks. Second, walled garden notifications work and are feasible, even though the actual usability of the notification and cleanup advice is currently rather poor. Third, since walled gardens are a recognized best practice for 'regular' botnet mitigation by ISPs, we can leverage the existing mitigation structures and practices to also help mitigate IoT botnets, rather than having to go through the time-consuming path of setting up new organizational structures and agreements.

There is a 'but', however. A significant one. The economic incentives for ISPs to adopt walled garden solutions are rather weak, as evidenced by the fact that only a fraction of the ISPs currently have them. Setting up and operating a walled garden, or operating any effective abuse management process in general, is a cost center for the ISP. Further eroding the incentives is the fear of customer pushback. Our analysis of customer experiences did indeed uncover a small but vocal minority that was angry or frustrated. Given the high cost of customer acquisition in these saturated markets, this fear might be enough to dissuade ISPs from quarantining infections. Overcoming this incentive problem might require a governmental measure to assign intermediate liability to ISPs. Soft versions hereof – e.g., a so-called 'duty of care' – already exist in many jurisdictions [110, 111].

While calling upon ISPs to take on this task, we can point out that their actions will have much higher chance of success than educating millions of end users about IoT security. Also, we can point to the fact a non-trivial portion of customers was pleased to be notified via the walled garden. As more people will become aware of the threats to their IoT devices, ISP mitigation might become more accepted – or even expected.

CHAPTER 7

Conclusion

This dissertation offers insights into the evidence-based design of security notification mechanisms to increase the effectiveness of voluntary action against cybercrime. It has introduced five peer-reviewed empirical studies that looked into the effectiveness of how vulnerable and abusive hosts and servers are remediated after communicating with the actors responsible for the affected systems or services. These studies were conducted to answer the following main research question:

How can the effectiveness of voluntary action against cybercrime be increased?

In the following section, we summarize the aims and outcomes of each study. This will be followed by lessons learned from the empirical studies, implications for governance of our findings and future work.

7.1 Summary of the Empirical Findings

Lack of evidence-based research on security notifications hindered the improvement of best practices on abuse and vulnerability handling and voluntary responses against cybersecurity problems.

Chapter 2 and Chapter 3 evaluated hosting providers' and domain owners' ability to remediate vulnerable and malicious servers. In these chapters, we focused on intermediary and direct remediation strategies. In **Chapter 2**, we measured the effectiveness of the abuse reports and the impact of the reputation of the abuse notification sender. The study used a private data feed of malware-infected websites to issue technically-similar abuse notifications from three senders with different reputations: an individual (low-reputation), a university (medium-reputation) and an established anti-malware organization (high-reputation). Our experiment results showed that 26% of the websites in the control group carried out a cleanup operation after 16 days, compared to 49%, 44%, and 48% for those that were assigned to the low-reputation, medium-reputation and high-reputation notification sender

groups respectively. Therefore, we found that our detailed abuse reports significantly increase the cleanup rates compared to not notifying. On the other hand, sender reputation did not significantly influence the cleanup process, as all notified groups with varying levels of reputation achieved similar cleanup rates. We also explored average cleanup rates of top 10 autonomous systems. We found a remarkably different cleanup rate for the various hosting providers which indicates that providers adopted different policies to deal with website infections.

Furthermore, we provided links to cleanup advice web pages to understand whether recipients can make use of these cleanup advice web pages. Surprisingly only around 7% of the resource owners and hosting providers visited our cleanup website. Nevertheless our results suggest providing a cleanup website containing specific instructions improves the cleanup speed when hosting providers view the instructions. Hosting providers that visited our cleanup websites had cleaned around 54% of the infected domains, while those who did not visit our cleanup websites had only cleaned around 29% of the infected websites after 3 days. On the other hand, this same positive impact is not shared by domain owners.

Following the experiment with malware-infected websites, **Chapter 3** investigated how underlying issues that cause compromises have been handled by the hosting providers and resource owners after reporting. Our study investigated the effectiveness of reaching out to different affected parties, and once reached incentivize for vulnerability remediation. We examine the effectiveness of reaching out through three notification channels: nameserver operators, domain owners, and network operators. The study compared the effectiveness of direct and intermediary remediation strategies in terms of remediation and reachability to find out which channel mobilizes the strongest incentive for remediation. We discovered that reaching out to both resource owners and intermediaries lead to more remediation than in the control group. On the other hand, the overall remediation rate remained very marginal. The majority of the servers remained vulnerable regardless of the remediation strategy we followed. Furthermore, our results found out that contacting the nameservers and domain owners at scale turned out to be an enormous problem. On average, 69% of notifications sent to nameserver operators via the SOA RNAME field generated a delivery failure. Moreover, email aliases we used for the second attempt generated even higher bounce rates than the SOA RNAME field. When we tried to reach resource owners via email addresses mentioned in domain WHOIS records, we observed that on average nearly 40% of our notification bounced. Meanwhile, notifications made to network operators bounced the least. On average, only 8% of our notifications made to network operators bounced. However, their remediation rates were slightly lower than the domain owner and name server operator groups. In the end, we concluded that there is no good communication mechanism for getting the wealth of vulnerability remediation information to the affected parties.

Additionally, we studied whether providing a link to a mechanism to verify the existence of the vulnerability could incentivize resource owners and intermediaries to act upon our notifications. Our results showed that only a few intermediaries and even fewer re-

source owners visited our websites. Only 12.2% of the name server operators and 14.75% of the network operators visited the demonstration website to test their nameservers. A comparison between remediation rates among treatment groups showed no evidence that vulnerability demonstrations did better than standard notification for both resource owners and intermediaries. However, the ones that visited the website achieved much higher remediation rates than the ones that received demonstrative notifications but did not visit the website. Similarly, our survey suggested that vulnerability demonstration was helpful and appreciated by the recipients.

Lastly, we investigated the reactions of the recipients through their email replies. In total, we have received 23 human replies and the majority of the notifications were positive. On the other hand, we received 5 negative replies and 2 neutral replies indicated that vulnerable servers did not belong to them.

In Chapter 2 and Chapter 3, we notified both owners of the insecure resources and intermediaries that facilitate the use of the infrastructure to host their resources. In Chapters 4, 5 and 6, we collaborated with an ISP to measure the effectiveness of notifications made to vulnerable and infected device owners. In these three chapters, we conducted experiments and collected empirical data on the remediation efforts of various broadband ISP subscribers. Additionally, we gained insights into their experience of how users perceive security notifications.

In **Chapter 4**, we studied user behavior and remediation effectiveness of walled garden notification on infected machines in a broadband ISP network. Our study describes the effectiveness of 1,736 quarantining actions involving 1,208 subscribers of a medium-sized ISP in the period of April-October 2017. We studied the relationship between cleanup rates and other factors, such as the release mechanism used to get out of quarantine, and the time spent in a quarantine environment. Our results illustrate that almost three-quarters of the quarantined users had managed to clean their infected machines in their first two attempts of quarantining when they have an option to self-release themselves from the quarantine environment. Significantly, providing an option to self-release from the quarantine environment did not introduce lax security behavior.

Moreover, we analyzed the quarantined user's communication with ISP employees to better understand their experience. We found that 27% of the quarantined subscribers request help from the employees of the ISP and 10% could not fix the problem and requested paid help. Additionally, we reported that around 10% of them voiced complaints about the disruption and about 3% even threatened to terminate their contract. Unfortunately, these numbers might discourage some ISPs to deploy the walled garden solution, as it represents an additional cost.

After disappointing remediation rates for large-scale vulnerability notifications in Chapter 3, we investigated the effectiveness of ISP issued email-only and walled garden vulnerability notifications compared to natural remediation. In **Chapter 5**, we measured the remediation rates achieved by a medium-sized ISP for 1,688 retail customers running open DNS resolvers or Multicast DNS services, which can be abused in amplification DDoS at-

tacks. We achieved a high remediation rate of around 87% for subscribers notified with walled garden notifications, even though quarantined subscribers can self-release from the quarantine environment without actually remediating the problem. Additionally, 75% of these subscribers that received email notifications managed to remediate. Remarkably, our study reported a high remediation rate for the control group: around half of all customers remediate, after excluding those customers who received notifications for different vulnerabilities. These observations have typically been removed from prior studies, which might explain the low remediation rates reported in those papers [33, 43].

After all, we analyzed communications logs between notified customers and the ISP employees to understand issues in vulnerability notification and remediation process. Interestingly, 16% of the notified users were unwilling to remediate because they did not wish to change the configuration of the device. They feared that the suggested remediation method will prevent them from using their devices and the functionality that comes with them. Similar to quarantined malware-infected machine owners, a high number of notified vulnerable device owners also requested additional help to solve the problem. Nearly half of all notified users contacted their ISP for additional information or help. Additionally, we found that about 11% of the notified subscribers complained about the disruptiveness of being quarantined. Of these almost 2% of them threatened to end their ISP subscription. Interestingly, these subscribers did not receive notifications consecutively as described in the previous study but still a high number of them contacted ISP for additional help and few of them even considered terminating their contact.

With the rise of IoT malware, cleaning up infected devices in ISP networks has become a critical task. In **Chapter 6**, we presented remediation rates from an observational study and a randomized controlled trial involving 220 consumers who suffered from Mirai infection. We found that improved walled garden notifications remediate 92% of the infections within 14 days. Similarly, walled garden notifications with standard notifications achieved high remediation rates, around 88%. Surprisingly, our results also demonstrated that the reinfection rate for standard walled garden notifications is low which suggests high long-term efficacy. On the other hand, email-only notifications have no observable impact compared to a control group where no notifications were sent. We measured high natural remediation rates of 74% for the control group. This could be due to the non-persistent nature of the IoT malware, which means turning off the device would get rid of the malware.

Finally, we investigated challenges in the cleanup process by analyzing user cleanup experience data collected via phone interviews and the communication logs of the ISP. Our results discovered that many users are operating from a wrong mental model, such as running antivirus software on their PC to solve IoT malware infection. Additionally, we found evidence that improving the content of the walled garden notification reduces the number of requests for additional help and complaints over the disruption of services. At the end of the study, the number of additional help requests and complaints were substantially lower in the improved walled garden group than the ones that were in standard walled garden group.

Our studies empirically measured the effectiveness of security notifications that drive voluntary action against cybercrime. The outcome of these studies provides a deeper and more detailed view of the effectiveness of different notifications mechanisms for vulnerabilities and abusive hosts in hosting and ISP networks. Combining insights on the user experience on remediation and effectiveness of different notification mechanisms, we contribute scientific evidence for establishing industry best practices around the voluntary remediation of vulnerable servers and devices and cleaning up abusive hosts. In the next sections, we revisit the lessons learned from empirical chapters of this dissertation.

7.2 Lessons learned

The model introduced in section 1.2 describes how abuse and vulnerabilities are reported to start the remediation process. This model guided our efforts to evaluate prior work and identify research gaps to better understand ways to improve the effectiveness of the abuse and vulnerability remediation process. Figure 7.1 illustrates lessons learned from each empirical chapter of this dissertation on abuse and vulnerability reporting infrastructure in eight different categories: (i) effectiveness of notifications; (ii) sender reputation; (iii) content; (iv) reachability; (v) reaction; (vi) usability; (vii) mechanism; and (viii) incentives. Each lesson will be explained in detail below.

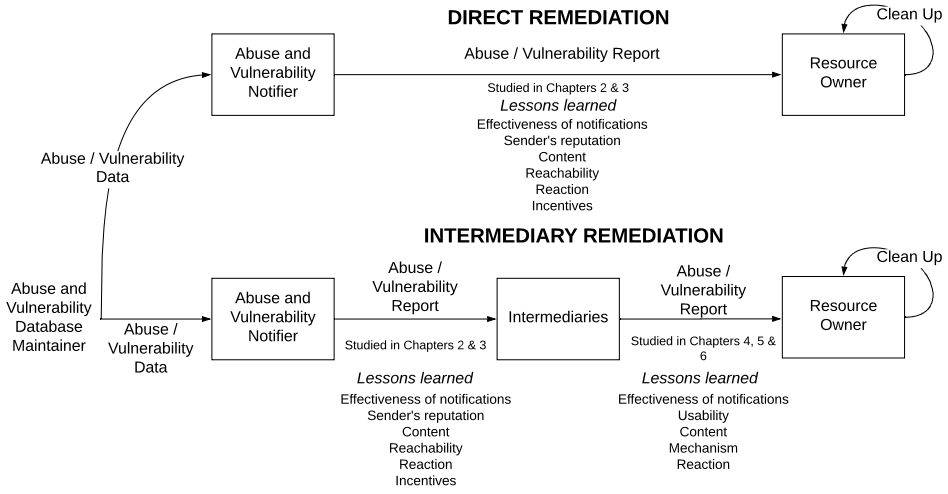


Figure 7.1: Aspects studied in this dissertation on abuse and vulnerability reporting infrastructure

7.2.1 Effectiveness of notifications

We demonstrated that our detailed abuse reports made from senders with varying reputations significantly increase compromised website cleanup compared to the control group. Interestingly, our results suggest a slower cleanup rate than reported by Vasek et al., who observed the majority of the infected websites notified with the same detailed abuse reports cleanup during 16 days experiment period [32]. On the other hand, they also reported a much higher cleanup rate for unnotified infected websites compared to our study. These might be because evasive techniques employed by Asprox slowed down the cleanup efforts more than common malware. Nonetheless, our results confirmed their findings on detailed abuse reports resulted in higher cleanup rates compared to not notifying. This shows that there is a surprising amount of voluntary action against malware-infected websites.

Furthermore, we found that vulnerability notifications made to both resource owners and intermediaries lead to more remediation than in the control groups. These results show that similar to abuse notifications, some entities acted upon our notifications and remediate the vulnerability. On the other hand, the overall vulnerability remediation rate remained marginal. Our results have been confirmed by both Stock et al. [44] and Li et al. [43] where vulnerability notifications made by these studies trigger more remediation than the control group, but overall remediation remained low.

All in all, these results confirmed that some intermediaries and resource owners voluntarily acted upon the notifications and remediated the security problems in question when they receive notifications from abuse and vulnerability reporters. Thus, sending security notifications increased remediation rates compared to not sending notifications. We recommend for abuse and vulnerability database maintainers and security companies to issue security reports regularly as these reports indeed increase the remediation. Noticeably, we have observed higher remediation rates for abuse notifications than vulnerability notifications made to the same actors in the hosting sector. This shows that abused or infected entities have stronger incentives to act upon the notifications than vulnerable but not abused parties. More details about incentives will be discussed in Section 7.2.8. That being said, we observed that the majority of the reports did not manage to trigger remediation for abuse or the vulnerability in question. Regardless, the conclusion is that we should look for alternative ways to incentivize more intermediaries and resource owners to act against cybercrime.

Our investigation in broadband ISP network found that both email and walled garden notifications made to vulnerable device owners achieved higher vulnerable device remediation than the control group. Slightly different results were reported while notifying infected IoT device owners. We found that only walled garden notification groups achieved higher remediation than not notifying. On the other hand, email-only notifications had no observable impact compared to not notifying. Thus, only walled garden notifications made a measurable difference compared to not notifying while remediating infected IoT devices.

Overall, our results in ISP networks demonstrated that nearly all infected IoT devices or vulnerable machines in ISP networks could be remediated via walled garden notifications

made by an ISP. Remarkably, the majority of the infected and vulnerable resources were remediated in the first walled garden attempt. This shows that with the right incentives and methods ISP can easily function as very effective control points to remediate vulnerabilities and abuse promptly. Having said that, email notifications only seem to make improvements on remediation when a security problem does not require more timely intervention. In brief, walled garden notifications are highly effective at increasing voluntary action against malware infections and vulnerabilities. On the other hand, while email notifications failed to increase voluntary action against infected IoT devices, they show promising results for less timely issues such as vulnerabilities. Thus, we suggest to use email notifications for vulnerabilities and less critical issues or create incentives schemes for email notifications to be more appealing for end users to act against cybercrime.

While we have demonstrated that notifications issued by ISP can be very effective at getting end users to remediate security issues, the cost of adoption and maintenance of these mechanisms by the ISPs remains a challenge. Future work will have to find methods to incentivize ISPs to participate in sending malware and vulnerability notifications to end users and join anti-botnet initiatives.

7.2.2 Sender reputation

Our results indicated no statistically significant difference between the treatment groups that belonged to the low-reputation, medium-reputation, and high-reputation senders. In the end, notifications from the anti-malware organization did not trigger more cleanup than notifications sent from an unknown individual researcher or university researchers. As a result of this, we demonstrated that sending abuse notifications from highly reputable senders does not always increase web-based malware cleanup. The only statistical difference found was between the control group and treatment groups, which is mentioned in the previous section. In short, the abuse notification sender's email address does not greatly matter when responding to abuse reports. However, sending abuse notifications, regardless of their sender's reputation, trigger more cleanup than not notifying.

One factor closely related to sender reputation that we did not investigate is the effectiveness of trusted entities. Typically, these entities would have a long-standing good relationship with the intermediary. As a result of this good relationship, the intermediary might be prioritizing notifications forwarded by trusted entities. It is also possible that notifications from trusted entities might not trigger higher cleanup rates due to higher friction in the cleanup process or lack of sender-based prioritizing by the intermediaries. In the end, future work will have to assess whether notifications from trusted entities can promote higher cleanup rates than others.

7.2.3 Content

Our results demonstrate that the majority of the recipients did not use our malware cleanup advice and vulnerability demonstration websites. Thus, we only assisted a few number of affected parties. This could be because links in email notifications trigger all kinds of overtones of phishing and drive-by download attacks. One simple possible approach to move forward might be to include the remediation advice in the notification content or to host these materials at trusted websites, such as the national CERT's website.

Additionally, we looked into the efficacy of websites created to assist the cleanup process. We found out that hosting providers that visited the cleanup web pages achieved higher remediation rate and speed than the ones that did not visit the pages. However, visiting the cleanup website did not make a difference in terms of remediation rates for the website owners. This suggests that basic cleanup advice can only enable hosting providers to achieve better cleanup. This might be because in general intermediaries have more experience in the remediation process and resource owners are technically less capable of following the advice. Potential improvements can be made on cleanup advice content to make it more actionable for the resource owners to cleanup web-based malware. Future work will have to investigate how to create such content and its effectiveness on remediation rates.

Prior research found that detailed abuse and vulnerability notifications trigger higher remediation rates than minimal notifications [43, 32]. In our notification studies, we adopted their detailed notifications and confirmed their conclusion on the effectiveness of detailed notifications compared to not notifying. In short, sending detailed notifications is an essential prerequisite for increasing voluntary action against web-based malware and vulnerabilities. On the other hand, we did not observe any improvements in remediation rates as we improve the content of the notifications by adding a mechanism to demonstrate vulnerability compared to standard detailed notifications. We don't know whether mechanism to demonstrate the vulnerability can be useful for other types of vulnerabilities. Perhaps, additional research can be conducted to evaluate our results on other types of vulnerabilities. Furthermore, significantly little research has been undertaken to evaluate the impact of offering cleanup tools for web-based malware cleanup. These tools have the potential to ease the cleanup by reducing the friction in the cleanup process. Similarly, we did not discover any significant improvements in remediation rates when we added more actionable content on the walled garden notifications for infected IoT device owners. That being said, both standard and improved walled garden notifications achieved outstanding rates of remediation and managed to remediate almost all infected IoT devices assigned to them. These rates are much higher than email notification with the same content as improved walled garden notifications. This makes us believe that the notification mechanism might have presented a more appealing incentive than the content of the notification. Crudely put, resource owners figure out a way to solve the security issue when the incentive is strong enough.

Additionally, our results confirmed that content has a major impact on the satisfaction

and ease of remediation. Reactions of the recipients suggest that improving the content reduces the fraction in the remediation process. For this reason, we encourage intermediaries and researchers to work further on how to create effective notification contents that can increase remediation rates as well as satisfaction levels of resource owners.

7.2.4 Reachability

Our results conveyed that nearly 80% of our email notifications bounced. These results demonstrated that delivering a security notification to both resource owners and intermediaries at scale is a significantly difficult problem. RFC standards, used for gathering contact details for security notifications, are only adopted by few resource owners and intermediaries. Only viable channel to report vulnerabilities and abusive behavior seems to be IP WHOIS abuse-email channel which is used by higher-level intermediaries. On the other hand, these email addresses typically belong to higher level intermediaries which are far removed from the resource or their operator. Thus, in many cases, they cannot do anything, except forwarding the notification to operators. In short, one of the major roadblocks to increase voluntary action against cybercrime is the reachability of correct contacts to remediate the security problems. Similar results have been mentioned by Stock et al. who also experienced significant reachability problems with their email vulnerability notifications [44].

We also investigated the reachability of email and walled garden notifications by conducting interviews with notified end users. The results of our interviews illustrated that even when private email addresses are used to contact the resource owners, there is a high chance that these notifications might have been overlooked or removed by the spam filters. We came across this after all the interviewees confirmed that the email accounts we used for the notifications were correct but, only half of them remember receiving the notification. One option to tackle this problem could be to maintain and regulate already existing notification channels. These channels can be monitored by higher intermediaries to make sure that email notifications can be delivered without failures to affected parties. Another option is to move away from email as the main notification medium to deliver notifications to intermediaries in the future. Intermediaries can be notified via more effective data sharing methods such as APIs or other types of data sharing methods. On the other hand, one possible issue could be the adoption of these data-sharing methods. Our studies show that in the current ecosystem intermediaries and resource owners cannot even set up working email addresses to receive abuse and vulnerability notifications. Therefore, it would be optimistic to imagine that these kinds of data sharing methods will be adopted by the owners of the resources and intermediaries without serious interventions. Lastly, intermediaries should also move away from email notifications to other types of warning systems to notify end users. For ISPs, walled garden notification systems seem to be a very promising tool to overcome the reachability problem. Similar systems can be deployed in the hosting market to notify domain owners and mitigate security problems if they refuse to

remediate the problem. Again, finding incentives and methods to promote the deployment of such notification and mitigation systems should be picked up by industry, governments, and researchers.

7.2.5 Usability

To evaluate the usability of the walled garden notifications, we looked into the relationship between remediation rates and usability related factors such as time spent in the quarantine environment and quarantine release mechanism. We pointed out that the majority of users spend a relatively short time in quarantine environments and still successfully removed the infection.

More interestingly, even the majority of the end users preferred the self-releasing option, we noticed that cleanup rates for this easy option and assisted release were very close together. The same positive behavior is observed while quarantining vulnerable device owners. Thus, we concluded that the self-releasing option does not introduce lax security behavior for both malware-infected machine owners as well as vulnerable device owners. The self-release option should be always present for initial quarantine events as it does not introduce lax security behavior and allows users to leave the quarantine environment at their convenience without any cost to ISP.

When we investigated release times of users after multiple events, we discovered that self-releasing users that managed to successfully clean up the malware infection left the walled garden faster over successive quarantine events. This makes us believe that there is a positive learning effect that promotes faster successful cleanup for users going through successive quarantining events. All in all, these results showed that the majority of the users managed to follow the instructions mentioned in walled garden content and remediate the problem in a relatively short time. Now, as a community, we know that basic remediation advice presented in walled garden notifications can help the majority of the notified subscribers to remediate security problems. In the future, we should test whether we can increase the effectiveness of voluntary action by producing more user-friendly notification contents. These types of notifications might incentivize less-skilled end users to take appropriate action.

7.2.6 Mechanism

Our investigation pointed out that both walled garden and email-only notification groups achieved much higher vulnerability remediation rates than the control group. Moreover, walled garden notifications promoted higher vulnerability remediation rates than email notifications even when notifications were made once a week. The same comparison on infected IoT device remediation demonstrated that walled garden notifications achieve significantly higher remediation rates than email-only notification mechanism while the email

group has no observable impact on infected IoT device remediation. By looking at the results, we illustrated that email notifications only make a difference in less critical issues where there is no active harm. That being said, user reactions demonstrated that the email-only notification mechanism triggers much fewer complaints and customer pushback than walled garden notifications.

All in all, we proved that walled garden notifications are very effective at getting resource owners to act against the security issue and triggers more remediation than email-only notification mechanism. Thus, using walled garden notifications increases the voluntary action against cybercrime much higher than email notifications. The main difference between walled garden and email notification mechanisms is the incentive walled garden provides. Email notifications can be easily ignored or overlooked, while the walled garden more forcefully urges users to read and act. Intermediaries should deploy and use walled garden notifications as it promotes more remediation. However, the walled garden also triggers more customer pushback than email notifications. Customer pushback can influence ISP's reputation and mainstream revenue as demonstrated by the complaints about the disruption and threats to terminate ISP subscriptions. On top of that, deploying and maintaining a walled garden system is a significant investment for an ISP due to providing additional help for end users in their attempts to remediate the security issues.

One way forward can be to look for ways to improve the content of walled garden notifications to decrease customer pushback and requests for additional help. As we introduced that more actionable walled garden notification content reduces the complaints and additional help requests. Furthermore, the effectiveness of alternative notification mechanisms and their cost-benefit analysis can be studied to find effective alternative notification mechanisms that trigger less pushback. Some researchers have pointed out that Instant Messenger (IM) could be an effective mechanism to notify end users [47]. Also, email notifications can be considered for less-critical security problems such as vulnerabilities as they trigger fewer complaints and more remediation than the control group.

7.2.7 Reaction

Our abuse reports for infected website owners and hosting providers did not trigger any negative reply. On the other hand, vulnerability notifications for hosting providers and domain owners trigger a few negative responses. A similar degree of negative replies has been mentioned in prior research [43]. These results demonstrate that recipients of our notifications are less familiar with vulnerability notifications and they did not understand the benefits of our scans and notifications. In short, replies towards our notifications were largely positive. Complementary to these findings, we found that user reactions are associated with cleanup actions. Our results indicated that human and automated responders achieved significantly higher cleanup rates than those that did not reply. In brief, proactive recipients tend to reply to the notification sender. This indicates that interaction may play a role in the successful cleanup.

Lastly, we observed reactions of end users notified via the walled garden and only-email notifications by their ISP. Our results showed that a large number of quarantined users with generic malware-infected machines contacted staff members for additional remediation related help and even some users requested a paid technician to solve the problem. Similar observations are made for quarantined users with vulnerable devices and even with infected IoT devices. This shows that notified end users request additional help to solve the problem regardless of the problem. That being said, we discovered that a significant number of these additional requests for help can be reduced by adding more actionable content to notifications.

Furthermore, we have noticed that a small group of end users contacted the ISP because they did not trust the authenticity of the walled garden or email notifications. Our interviews with IoT infected device owners pointed out that improving the content of the notification by adding actionable content increases trust towards ISP-made security notifications even when the notification is not personalized. Once again, this displays the importance of containing information that allows non-expert actors to reliably tell the notifications apart from a random phishing page. In short, notification senders should consider personalizing the notification and adding more actionable content to increase the credibility of the notifications.

Moreover, we observed that a fraction of users voiced complaints about the disruption and even threatened to terminate their contract. These negative reactions are given mostly because of quarantine. Email notifications did not trigger negative reactions. We observed similar complaints voiced by quarantined vulnerable device owners. Again, more actionable notification content proven to be reducing the percentage of complaints.

All in all, these results showed that quarantining infected and vulnerable device owners is a disruptive treatment and triggers more remediation as well as negative reactions than email notifications. One way to reduce complaints, trust issues, and additional help requests is to improve the notification content by providing more actionable and personalized content. In the future, intermediaries and researchers should work together to create more effective notification content to solve these types of problems.

7.2.8 Incentives

Throughout this dissertation, we looked into different incentive mechanisms to understand and improve current abuse and vulnerability remediation processes. Initially, we assessed the incentives of resource owners and intermediaries to understand whether it was more effective to leverage direct or intermediary remediation strategies to remediate the vulnerability. Our results demonstrated that intermediaries and resource owners have equally weak incentives to act upon the notification and remediate the vulnerability. In the end, both direct and intermediary remediation strategies displayed significant reachability and remediation challenges. Hence, the majority of the servers remained vulnerable. We thought that intermediaries might be reluctant to act because resources affected by the zone poisoning

attack do not belong to the intermediaries. Thus, intermediaries will not suffer the full cost of failure when vulnerable resources are compromised. Meanwhile, resource owners might be reluctant to act because of the high friction in the process towards remediation. In any case, we could not create any incentive structure that could make a significant difference in terms of remediation. Since we could not identify a clear entity to direct our notifications, we suggest notifying all entities responsible for remediating the vulnerable system or service to increase remediation rates. Another approach could be to deploy reputation effects (naming, praising and shaming) to increase vulnerability remediation rates for the providers. Reputation effects are considered as an alternative type of regulatory intervention [112]. This intervention mechanism might increase remediation rates over a longer time frame as demonstrated by prior research [37].

Moreover, we investigated incentives to remediate abuse compared to vulnerability. The outcome of our notification experiments showed that abuse reports trigger more remediation than vulnerability notifications. Similar observations can be seen by looking into previous research. For example, detailed reports from Vasek et al. remediated the majority of malware-infected websites while remediation rates for vulnerabilities were low even when other mechanisms of notifications are used to notify the affected parties [45, 44, 43]. These results suggest that hosting providers and domain owners prioritize abuse over vulnerability notifications. There are few exceptions to this situation where vulnerabilities are heavily publicized. For instance, notifications for high-profile disclosure of the Heartbleed vulnerability reached significantly higher remediation than the control group [41]. This shows that the effectiveness of the vulnerability notification campaigns can be enhanced by publicizing the vulnerability.

Lastly, we studied incentives of walled garden notification mechanisms for remediation compared to the email notification mechanism. Our experiment results found that walled garden notification mechanism leads to better vulnerability and IoT malware infection remediation compared to email notifications and not sending notifications. This is because the walled garden provides more incentives for remediation than email notifications. Consequences for ignoring walled garden notification will be placed back to the quarantine environment where only limited access to the Internet is allowed. On the other hand, ignoring email notifications could trigger more email notifications. Preventing access to the Internet is an effective incentive mechanism to increase remediation. Unfortunately, this type of alternative incentive mechanisms also increases complaints and cost to ISP-made notification systems. In short, sentiments of reactions and remediation rates depend on the incentives of the notification mechanism. In the future, industry and academics should focus on finding ways to reduce the complaints and cost associated with the walled garden notifications. Alternatively, the cost and effectiveness of other notification mechanisms should be investigated to look for effective notification mechanisms that present less cost.

7.3 Implications for Governance

Most of what is being done to remediate abusive and vulnerable hosts is carried out by private actors. Even if these actors are not legally required to act against security problems, they do. Our research provides a better understanding of how effective these actors are in terms of abuse and vulnerability remediation and how can they be more effective in hosting and ISP market. To answer these questions, we assessed the effectiveness of abuse and vulnerability notifications and identified factors that make voluntary regimes more effective, both in terms of design and incentives.

This section offers implications for governance of this research. In order to address complex issues, governance literature points out four canonical modes of governance: (i) market; (ii) hierarchy; (iii) network and; (iv) community [113, 114]. We provide various instruments from these four modes to address issues related to reachability and incentives for remediation in hosting and ISP markets. We focused on these two areas because they require interventions at a higher level than the firm level. Collaboration between firms, actors, and stakeholders are required to improve reachability and incentives for remediation. On the other hand, other issues mentioned in the previous section can be addressed at the individual firm level because firms can implement these themselves. For example, intermediaries can decide which mechanisms or content to use while notifying malware-infected resource owners.

7.3.1 Increasing the reachability

A crucial challenge that we ran into in our notification experiment is to find a contact point for vulnerable or abused resources. This issue has significantly undermined the effectiveness of abuse and vulnerability notification campaigns on a global scale. We encountered two governance structures that were directly involved in reachability issues. The first one is the lack of RFC-compliant email address adoption among resource owners and the second one is WHOIS records being outdated, wrong or missing certain information. Here we are going to focus on ways to improve inaccurate WHOIS records, because contact points in WHOIS records bounced less than emails sent to RFC-compliant email addresses. These contact points are regulated through a series of intermediaries that can be influenced by governance mechanisms.

During the domain registration process, the domain registrant required to provide accurate contact information to the domain registrar. Once contact information is provided, the domain registrar is responsible for updating the WHOIS Domain database with this information. WHOIS Domain database is filled and managed by domain registrars and regulated by Internet Corporation for Assigned Names and Numbers (ICANN). Moreover, ICANN outlined a set of rules with registrars to provide and maintain up-to-date contact information. Registrar can maintain this by making quarterly validation checks to assess the correctness of the information. If registrants do not want to provide up-to-date contact

information, then the domain can be suspended by the registrar to satisfy the rules outlined by ICANN.

IP registration follows a different process. ICANN allocates IP addresses to five Regional Internet Registries (RIRs). From there on, these five RIRs further allocate IP addresses to hosting providers and ISPs. Lastly, providers sub-allocate IP addresses to organizations and end users. RIRs are responsible for maintaining the WHOIS IP database. For instance, RIPE NCC, one of the five Regional Internet Registries (RIRs), oversees the IP address allocations and registrations for Europe, West Asia, and Russia and maintains WHOIS records for IP addresses allocated to these regions. Below we will discuss governance mechanisms to regulate the accuracy of contact information in the WHOIS databases.

To address reachability issues, we highlighted in our empirical studies, a combination of governance mechanisms mentioned below.

Market

This mode of governance relies on contracts between parties, efficient resource use, and market competition. Our results demonstrated that finding the correct contact point for affected parties is a very challenging issue. After all, lack of incentives and cost associated with improvements are causing resource owners and intermediaries to push inaccurate and incorrect contact information in the WHOIS Domain records. These records should be detected and improved as soon as possible but the cost associated with maintenance reduces involved actors' willingness to act. Yet, typically benefits of maintaining the accuracy of contact points for abuse and vulnerability notifications go to all intermediaries and resource owners. This can be seen as a market failure because intermediaries that maintain correct contact points of themselves and their customer incurs the cost of maintenance, while the security benefits diffuse to society.

ICANN is already aware of the situation and has addressed with few implementations on their own. One of them was to introduce the ICANN-Accredited Registrars list [115]. This list contains the names of domain registrars that have the Registrar Accreditation Agreement 2013 (RAA 2013) with ICANN to act as a domain registrar for generic top-level domains (gTLDs) [66]. This put a contract in place between ICANN and registrars. In the context of this contract, ICANN requires registrars to maintain the integrity of the WHOIS database [116]. On the other hand, not all registrars are actively checking whether data provided in WHOIS is correct. This causes WHOIS data to be inaccurate. One way to enforce this is through price regulations. ICANN can force domain registrars with a lack of WHOIS data accuracy control to pay more for each domain registration. This type of cost could incentivize registrars to place basic checks in place. Alternatively, if a domain registrar knew to lack any mechanism to validate contact details, ICANN can incentivize registrars by suspending their ability to register domains. This approach would be the ultimate incentive to enforce registrars to provide accurate WHOIS database.

Hierarchy

Typically, hierarchical governance refers to making laws, legislation, and regulations to regulate markets or organizations. Market failures can be addressed through laws and regulations. State-specific instruments might have a limited impact on increasing the reachability of the affected parties. This is because online resources and intermediaries are globally distributed which makes them difficult to govern through national laws.

Nonetheless, hierarchy governance could be used to address reachability issues in the hosting sector geographically. For example, governments at national levels could make all nationally registered companies that have registered domains and IP ranges to provide a functioning contact point to remediate abuse and vulnerabilities in the WHOIS database. This way at least one contact point for both resource owners and intermediaries will be collected. Later on, these contact points can be shared with local CERTs or abuse notifiers via WHOIS records to disseminate abuse or vulnerability data.

Fines can be leveraged by governments to incentivize nationally registered companies, providing valid contact points for abuse and vulnerability reporting. Companies that fail to provide or maintain their contact point for abuse and vulnerability reporting could be fined. These types of measures could generate economic incentives for registered companies.

Network

Network governance can be seen as continuous collaborative interactions among independent companies and organizations. It could be based on trust, shared goals or resources. This mode of governance can be used to address reachability issues through collaborations in public-private partnerships. On the other hand, it could be challenging to scale and maintain such networks in the hosting and ISP sectors which are distributed around the globe. For this reason, focusing on broader audiences could significantly reduce the reachability issues in abuse and vulnerability reporting. For instance, at the EU level, where the European Commission would work with RIPE NCC and other Regional Internet Registries (RIRs) to create an initiative to assess and enforce the correctness of abuse contact information in the WHOIS.

Alternatively, the same initiative could also participate in conducting awareness campaigns for domain owners to adopt RFC-compliant email boxes that can be used to deliver abuse and vulnerability notifications.

Community

Typically, communities consist of various organizations and individuals with similar characteristics and identity. They share and promote joint norms and values. Communities play a significant role in the governance of the hosting market. They develop and share best practices for abuse and vulnerability mitigation and remediation. That's why community organizations can be effective in promoting strategies to increase voluntary actions against

cybercriminal activities in hosting services. An example of this is Messaging, Malware and Mobile Anti-Abuse Working Group (M3AAWG), an industry organization with network operators as members. Members come together to figure out ways of handling abuse and vulnerabilities in their network. Our results can be used to identify areas to improve and develop best practices to support the community members. Additionally, on a general level, our experiment results can be used as general guidelines by security companies and other abuse and vulnerability notifiers for avoiding certain pitfalls in abuse and vulnerability reporting.

7.3.2 Improving the incentives

Providing strong incentives for cleanup of more generic compromised resources and the remediation of vulnerabilities is one of the most significant components of the fight against cybercrime. Here, we discuss governance instruments that are worth considering to increase the effectiveness of voluntary defense mechanisms.

Market

Our findings demonstrate that some intermediaries and resource owners voluntarily act on abusive and vulnerable resources in their networks although damage is an externality to them. On the other hand, the extent of this action varies strongly and changes regularly based on characteristics of the notification as seen by varying abuse and vulnerability remediation rates. Moreover, acting against abuse and vulnerabilities incurs a cost while the benefits of acting are typically shared by society. Therefore, it is quite possible to see failures that can undermine the effectiveness of voluntary action.

Our notifications were made privately. We did not publicly display or rank the provider's ability to remediate the security problems mentioned in the notifications. Besides, we did not know whether our notifications reached the affected party nor what kind of response was given to our notification. Similarly, other abuse and vulnerability reporters also experience the same situation. All this presents a remediation ecosystem with limited transparency. Lack of transparency might cause providers to ignore security notifications to avoid costs associated with the remediation. Moreover, server and website owners would not know whether their provider invests enough resources to address these problems. They were also unaware of the consequences associated with a lack of response to abuse. One way to increase the effectiveness of voluntary action against cybercrime could be by making abuse and vulnerability notifications and their responses publicly available to all intermediaries and resource owners. In this instance, the incentive mechanism would be reputational. Intermediaries that are not acting upon the notifications could receive pressure from their local CERTs, other intermediaries and their customers. This could incentivize intermediaries to invest more into abuse and vulnerability remediation. Work from Tang et al. on publicizing spam ranking per provider suggests that such an approach could be very effec-

tive at reducing security issues[37]. This approach would provide an economic incentive for hosting providers to invest more in remediation because security concerns customers might avoid hosting providers who ignore notifications.

Hierarchy

Issues in abuse and vulnerability remediation are mainly due to the difference in interests among public and private actors. The main aim of private actors is to make a profit. As a result of this, they invest less than social optimum to remediate vulnerabilities and abusive sources in their network because of the cost associated with remediation being an externality to them. Meanwhile, the public sector concerned with reducing cybercrime.

Hierarchical governance mechanisms could play a significant role here by providing financial incentives to increase providers' willingness to adopt abuse sharing platform and act on security notifications. This can be done through tax law or subsidies. One example could be to allow intermediaries and other companies to deduct the cost of joining an abuse sharing platform from their profits so that they invest more in remediation and mitigation. Alternatively, tax cuts and subsidies can be provided to those who regularly act upon the security notifications at a satisfactory level or adopt effective security mechanisms such as walled garden notifications. This mechanism requires yearly assessments to evaluate the security levels of the providers. The same assessment can be used to improve information transparency. Based on this, various other governmental interventions can be taken to further incentivize to increase providers' willingness to act. In addition to rewarding, another instrument could be to penalize providers with unsatisfactory levels of abuse and vulnerabilities in their network through fines. Recently a Dutch law is being drafted to fine hosting providers with a lax attitude towards child pornography on their servers could be a good example to penalize lax security efforts through fines[117]. According to this law, providers have to remove such material within 24 hours to prevent any kind of penalties. Proposed time interval for removal suggested by industry. Thus, this can be also seen as a constructive interaction between hierarchical and network governance.

Network

Our results show that intermediaries can remediate the majority of the abuse and vulnerabilities in their network once they are reported. However, mechanisms we tested result in a significant amount of additional help requests and customer push back. All this increases the cost of voluntary action against cybercrime. After all, the cost associated with these systems and customers pushes back might decrease providers' willingness to adopt effective security measures.

That being said, long-term relationships can be leveraged to promote the adoption of effective notification mechanisms or reduce the cost of acting against cybercrime for intermediaries and resource owners. One way to accomplish this is through a public-private

partnership. For instance, national institutions could collaborate with intermediaries to create initiatives to promote public-private clearinghouse membership for abuse data and effective notification mechanisms. One example of the public-private clearinghouse is AbuseHub. The initiative is created to provide more accurate abuse and vulnerability data for Dutch ISPs. While the development cost of this initiative funded jointly by both government and ISPs, maintenance cost covered by solely ISPs. This helped reduce the cost of acting against abuse by all of the member ISPs. Additionally, this kind of initiative would help accomplish a set of rules and agreed norms on abuse and vulnerability remediation. Evidence from public AbuseHUB reports suggests that Dutch ISPs improved substantially in terms of botnet mitigation[118].

Additionally, all the remediation advice and tools for end users can be collected on one trusted site to be used by all intermediaries and resource owners. This site can be hosted by governmental entities but the content of the website can be produced by intermediaries based on their experience with their customers. Intermediaries can refer to this site when they notify their customers or promote awareness campaigns for common security practices. Thus, this would reduce intermediaries' cost of giving cleanup advice and liability for the harm when cleanup advice fails to mitigate the problem. One example of this can be a Dutch security advice website called veiliginternetten.nl. The website provides tips on how to stay safe and secure while browsing the Internet. This site is an initiative of the Ministry of Economic Affairs and Climate and the Ministry of Justice and Security and many private entities.

Community

Security communities are very proactive at providing best practices for abuse and vulnerability remediation. However, the effectiveness of these documents was never studied. Furthermore, best practices are not widely adopted by the community. Our studies helped hosting and ISP communities in two unique ways: (i) providing empirical evidence on how to increase abuse and vulnerability remediation; and (ii) identifying new areas to address. Providing the effectiveness of security mechanisms could incentivize community members to adopt these mechanisms more widely. Our results could help to establish new norms to increase the effectiveness of voluntary action against cybercrime among community members. For example, our results showed that walled garden notifications promote significantly higher cleanup rates than email notifications. This could incentivize providers to adopt walled garden notifications more widely. Alternatively, we identified proactive providers that act upon abuse and vulnerability notifications in a timely manner. This can be used to signal the social norms to the less active providers.

Lastly, our studies highlighted which mechanisms and factors are effective at getting users to act against abuse and vulnerabilities in hosting and ISP markets. These findings support the evidence-based design of new best practices to increase the effectiveness of voluntary action against cybercrime.

7.4 Limitations and Future Work

Key limitations and future work of each empirical study in this dissertation described in its chapter. Here, we discuss 3 future directions and limitations that were encountered in our empirical and prior studies.

First limitation is the generalizability of our results to other intermediaries, notification mechanisms or remediation of other security problems. In each empirical study, we use data sources available to us to conduct our experiment and draw conclusions. Typically, these chapters study the impact of various notifications mechanisms on a single type of security problem or towards those that are sinkholed. Thus, we recommend conducting follow-up studies to understand the generalizability of our treatments on other types of infections and vulnerabilities. Additionally, conclusions for notification studies in the ISP network are tied to data from a single ISP in Europe. We observed different natural remediation rates across different networks in the same ISP. Thus, it is hard to generalize our results for other ISPs. In the future, studies can be extended by outcomes of the notifications from other ISPs with different characteristics and jurisdictions. Therefore, the generalizability and reproducibility of our results to other ISPs or networks are a matter for further research.

Another limitation encountered by all notification studies is having a lack of visibility into how notifications are processed, understood and acted upon by the affected parties. Typically, researchers receive very few acknowledgments. In many cases, no acknowledgement is made by the affected parties. This makes it difficult to understand the successful delivery of the notification and response given. Similarly, in many notification studies, the impact of vulnerability and abuse notifications from other sources are either ignored or overlooked. In the empirical chapters that we collaborated with an ISP, we had more visibility into the remediation process but, we lack insights into how remediation is performed and how resources are naturally remediated without any notification from ISPs. Asking resource owners what they did to remediate the security problem is a difficult problem because they might not remember what they did. Visibility into how remediation is performed remains a challenge that has to be addressed by the security community, including researchers, network providers, and CERTs.

Lastly, we lack insights into how notified resource owners improved over time. Prior studies did not look into long term effects of abuse and vulnerability notifications in the hosting market. Thus, as a community, we lack remarkable insights into the kinds of notification methods that can provide long term security benefits to server and domain owners. Such research would help the security community to identify the most effective methods to prevent revictimization and even guidance for faster recovery. Similarly, we did not look far into how quarantined users keep their resources secure. We could not study the effects of such intervention because of partnered ISP's storage constraints. However, collecting long term effects of the walled garden and other types of notifications could help us understand the broader effects of ISP-made notifications. We hope that future work on notifications can investigate long term effects of notifications and look for kinds of notifications that can

incentivize for long term security benefits.

Bibliography

[1] N. Woolf, "DDoS attack that disrupted internet was largest of its kind in history," https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet, 2016.

[2] Cloudflare, "What is the Mirai Botnet?" https://www.cloudflare.com/learning/ddos/glossary/mirai-botnet/, 2019.

[3] Graham Cluley, "These 60 dumb passwords can hijack over 500,000 IoT devices into the Mirai botnet," 2016, https://www.grahamcluley.com/mirai-botnet-password/.

[4] T. Claburn, "Mirai bots' cyber-blitz 1m german broadband routers – and your isp could be next," https://www.theregister.co.uk/2016/11/28/router_flaw_exploited_in_massive_attack/, 2016.

[5] B. Krebs, "New Mirai Worm Knocks 900K Germans Offline," https://krebsonsecurity.com/2016/11/new-mirai-worm-knocks-900k-germans-offline/, 2016.

[6] J. Leyden, "How hack on 10,000 WordPress sites was used to launch an epic malvertising campaign," https://www.theregister.co.uk/2018/07/30/malvertising_wordpress/, 2018.

[7] Ionut Ilascu, "25 Million Android Devices Infected by Agent Smith Malware," 2019, https://www.bleepingcomputer.com/news/security/25-million-android-devices-infected-by-agent-smith-malware/.

[8] K. Townsend, "18.5 million websites infected with malware at any time," https://www.securityweek.com/185-million-websites-infected-malware-any-time, 2018.

[9] C. Cimpanu, "Z-Shave Attack Could Impact Over 100 Million IoT Devices," https://www.bleepingcomputer.com/news/security/z-shave-attack-could-impact-over-100-million-iot-devices/, 2018.

[10] C. Edwards, "More than a million asus laptops have been hacked – how to check if you're affected," https://www.thesun.co.uk/tech/8722161/asus-laptops-hacked-malware-million-check-if-affected/, 2019.

[11] National Vulnerability Database, 2019, https://nvd.nist.gov/.

[12] X. Ugarte-Pedrero, M. Graziano, and D. Balzarotti, "A close look at a daily dataset of malware samples," *ACM Transactions on Privacy and Security (TOPS)*, vol. 22, no. 1, p. 6, 2019.

[13] Shodan, 2019, https://shodan.io.

[14] The ZMap Project, 2019, https://zmap.io/.

[15] D. Miessler, "Masscan Examples: From Installation to Everyday Use," https://danielmiessler.com/study/masscan/, 2018.

[16] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A Search Engine Backed by Internet-Wide Scanning," in *22nd ACM Conference on Computer and Communications Security*, Oct. 2015.

[17] C. Osborne, "Most enterprise vulnerabilities remain unpatched a month after discovery," https://www.zdnet.com/article/the-majority-of-vulnerabilities-remain-unpatched-a-month-after-discovery/, 2018.

[18] M. Korolov, "Majority of websites have serious, unfixed vulnerabilities," https://www.csoonline.com/article/2928006/majority-of-websites-have-serious-unfixed-vulnerabilities.html, 2015.

[19] D. Oreilly, "Detect and prevent today's sophisticated malware threats," https://www.cnet.com/how-to/detect-and-prevent-todays-sophisticated-malware-threats/, 2012.

[20] "Conficker Working Group," http://www.confickerworkinggroup.org.

[21] M. H. Jhaveri, O. Cetin, C. Gañán, T. Moore, and M. V. Eeten, "Abuse reporting and the fight against cybercrime," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, p. 68, 2017.

[22] PhishTank, 2019, https://www.phishtank.com/.

[23] L. Constantin, "Google safe browsing alerts network admins about malware distribution domains," https://www.pcworld.com/article/245373/google_safe_browsing_alerts_network_admins_about_malware_distribution_domains.html, 2011.

[24] Shadowserver, 2019, https://www.shadowserver.org/.

[25] L. H. Newman, "How Google's Safe Browsing helped build a more secure web," https://www.wired.com/story/google-safe-browsing-oral-history/, 2018.

[26] "Friends of PhishTank," http://phishtank.com/friends.php, 2019.

[27] "Avoid and report phishing emails," https://support.google.com/mail/answer/8253?hl=en, 2019.

[28] M. Van Eeten, J. Bauer, H. Asghari, S. Tabatabaie, and D. Rand, "The role of internet service providers in botnet mitigation an empirical analysis based on spam data," in *Research Conference on Communications, Information and Internet Policy (TPRC)*, 2010.

[29] T. Casey, "Survey: 27 percent of it professionals receive more than 1 million security alerts daily," https://www.imperva.com/blog/27-percent-of-it-professionals-receive-more-than-1-million-security-alerts-daily/, 2018.

[30] "Network reporting," https://www.shadowserver.org/what-we-do/, 2019.

[31] Robert Abel, "Github announces 4 million vulnerabilities patched in half a million repositories," 2018, https://www.scmagazine.com/home/network-security/github-announces-4-million-vulnerabilities-patched-in-half-a-million-repositories/.

[32] M. Vasek and T. Moore, "Do malware reports expedite cleanup? an experimental study," in *5th Workshop on Cyber Security Experimentation and Test, CSET*, 2012.

[33] F. Li, G. Ho, E. Kuan, Y. Niu, L. Ballard, K. Thomas, E. Bursztein, and V. Paxson, "Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension," in *Proceedings of the 25th Int. Conference on WWW*, 2016, pp. 1009–1019.

[34] D. Canali, D. Balzarotti, and A. Francillon, "The role of web hosting providers in detecting compromised websites," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 177–188.

[35] A. Nappa, M. Z. Rafique, and J. Caballero, "Driving in the Cloud: An Analysis of Drive-by Download Operations and Abuse Reporting," in *Proceedings of the 10th Conference on Detection of Intrusions and Malware & Vulnerability Assessment*. Berlin, Germany: Springer, July 2013, pp. 1–20.

[36] C. Gañán, O. Cetin, and M. van Eeten, "An Empirical Analysis of ZeuS C&C Lifetime," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*. ACM, 2015, pp. 97–108.

[37] Q. Tang, L. Linden, J. S. Quarterman, and A. B. Whinston, "Improving internet security through social information and social comparison: A field quasi-experiment," *WEIS 2013*, 2013.

[38] M. Vasek, M. Weeden, and T. Moore, "Measuring the impact of sharing abuse data with web hosting providers," in *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security*. ACM, 2016, pp. 71–80.

[39] T. Moore and R. Clayton, "The consequence of non-cooperation in the fight against phishing," in *2008 eCrime Researchers Summit*. IEEE, 2008, pp. 1–14.

[40] A. Hutchings, R. Clayton, and R. Anderson, "Taking down websites to prevent crime," in *Electronic Crime Research (eCrime), 2016 APWG Symposium on*. IEEE, 2016, pp. 1–10.

[41] Z. Durumeric, J. Kasten, D. Adrian, J. A. Halderman, M. Bailey, F. Li, N. Weaver, J. Amann, J. Beekman, M. Payer *et al.*, "The Matter of Heartbleed," in *Proceedings of the 2014 Conference on IMC*. ACM, 2014, pp. 475–488.

[42] M. Kührer, T. Hupperich, C. Rossow, and T. Holz, "Exit from Hell? Reducing the Impact of Amplification DDoS Attacks," in *USENIX Security Symposium*, 2014.

[43] F. Li, Z. Durumeric, J. Czyz, M. Karami, M. Bailey, D. McCoy, S. Savage, and V. Paxson, "You've got vulnerability: Exploring effective vulnerability notifications," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016, pp. 1033–1050.

[44] B. Stock, G. Pellegrino, C. Rossow, M. Johns, and M. Backes, "Hey, you have a problem: On the feasibility of large-scale web vulnerability notification," in *USENIX Security Symposium (Aug. 2016)*, 2016.

[45] B. Stock, G. Pellegrino, F. Li, M. Backes, and C. Rossow, "Didn't You Hear Me?–Towards More Successful Web Vulnerability Notifications," in *The Network and Distributed System Security Symposium (NDSS)*, 2018.

[46] E. Zeng, F. Li, E. Stark, A. P. Felt, and P. Tabriz, "Fixing HTTPS Misconfigurations at Scale: An Experiment with Security Notifications," in *The 2019 Workshop on the Economics of Information Security (WEIS 2019)*, 2019.

[47] J. Zhang, H. Duan, W. Liu, and X. Yao, "How to Notify a Vulnerability to the Right Person? Case Study: In an ISP Scope," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–7.

[48] T. Moore, R. Clayton, and R. Anderson, "The economics of online crime," *Journal of Economic Perspectives*, vol. 23, no. 3, pp. 3–20, 2009.

[49] T. Moore and R. Clayton, "Examining the impact of website take-down on phishing," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. ACM, 2007, pp. 1–13.

[50] D. Crocker, "Mailbox Names for Common Services, Roles and Functions," RFC 2142 (Proposed Standard), Internet Engineering Task Force, May 1997. [Online]. Available: http://www.ietf.org/rfc/rfc2142.txt

[51] StopBadware, 2019, http://www.stopbadware.org/.

[52] E. Schwartz, "The moz blog study: How searchers perceive country code top-level domains," 2014, http://moz.com/blog/cc-tld-domain-study.

[53] StopBadware, "Best Practices for Reporting Badware URLs," 2011, https://www.stopbadware.org/files/best-practices-for-reporting-badware-urls.pdf.

[54] J. d. T. Nart Villeneuve and D. Sancho, "Asprox Reborn," Trend Micro Incorporated, Tech. Rep., 2009.

[55] VirusTotal, "Searching with VirusTotal," 2015, https://www.virustotal.com/en/documentation/searching/#getting-url-scans.

[56] Sucuri Malware Labs, "Php error: Fatal error," 2015, http://labs.sucuri.net/db/malware/php-error-fatal-error?v6.

[57] SEO Tools, "Server header checker," 2019, http://tools.seobook.com/server-header-checker.

[58] Cloudflare, "Content delivery network," 2019, https://www.cloudflare.com/cdn/.

[59] National Institute of Standards and Technology, "StopBadware Commentary on Liability of Web Hosts for Malware Distribution," 2011, http://www.nist.gov/itl/upload/StopBadware_Web-Hosting-Provider-Liability-for-Malicious-Content.pdf.

[60] T. Moore and R. Clayton, "Evil searching: Compromise and recompromise of internet hosts for phishing," in *Financial Cryptography and Data Security*, ser. Lecture Notes in Computer Science, R. Dingledine and P. Golle, Eds., vol. 5628. Springer, 2009, pp. 256–272.

[61] O. Cetin, M. Hanif Jhaveri, C. Gañán, M. van Eeten, and T. Moore, "Understanding the role of sender reputation in abuse reporting and cleanup," *Journal of Cybersecurity*, vol. 2, no. 1, pp. 83–98, 2016.

[62] M. Korczyński, M. Krol, and M. van Eeten, "Zone Poisoning: The How and Where of Non-Secure DNS Dynamic Updates," in *Proceedings of the 2016 ACM on Internet Measurement Conference*. ACM, 2016, pp. 271–278.

[63] P. Vixie, S. Thomson, Y. Rekhter, and J. Bound, "Dynamic Updates in the Domain Name System (DNS UPDATE)," Internet RFC 2136, April 1997.

[64] "Abuse Contact DB," accessed: 2017-02-21. [Online]. Available: https://www.abusix.com/contactdb

[65] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: an empirical study of the effectiveness of web browser phishing warnings," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 1065–1074.

[66] ICANN, "Registrar Accreditation Agreement," 2013. [Online]. Available: https://www.icann.org/resources/pages/approved-with-specs-2013-09-17-en#privacy-proxy.2.1

[67] "A list of free email provider domains," 2019. [Online]. Available: https://gist.github.com/tbrianjones/5992856

[68] S. He, G. M. Lee, S. Han, and A. B. Whinston, "How would information disclosure influence organizations' outbound spam volume? evidence from a field experiment," *Journal of Cybersecurity*, 2016.

[69] J. Livingood, N. Mody, and M. O'Reirdan, "Recommendations for the Remediation of Bots in ISP Networks (RFC 6561)," *Internet Eng. Task Force*, 2012.

[70] Messaging Anti-Abuse Working Group and others, "Abuse Desk Common Practices," 2007. [Online]. Available: https://www.m3aawg.org/sites/default/files/document/MAAWG_Abuse_Desk_Common_Practices.pdf

[71] U. Jilani, "The ACMA and Internet providers working together to combat malware," 2015. [Online]. Available: https://www.acma.gov.au/theACMA/engage-blogs/engage-blogs/Cybersecurity/The-ACMA-and-internet-providers-working-together-to-combat-malware

[72] ECO Internet industry association, "Botfree," 2013. [Online]. Available: https://www.botfree.eu/en/aboutus/information.html

[73] International Telecommunication Union (ITU), "ITU Botnet Mitigation Toolkit," https://www.itu.int/ITU-D/cyb/cybersecurity/projects/botnet.html, 2018.

[74] European Network and Information Security Agency (ENISA), "Involving Intermediaries in Cyber-security Awareness Raising," 2012. [Online]. Available: https://www.enisa.europa.eu/publications/involving-intermediaries-in-cyber-security-awareness-raising

[75] National Institute of Standards and Technology, "Models To Advance Voluntary Corporate Notification to Consumers Regarding the Illicit Use of Computer Equipment by Botnets and Related Malware," 2011. [Online]. Available: https://www.nist.gov/itl/upload/SANS_BotNet-FRN-Comment-11-4-11.pdf

[76] H. Asghari, M. Ciere, and M. J. Van Eeten, "Post-mortem of a zombie: conficker cleanup after six years," in *USENIX Security Symposium*. USENIX Association, 2015, pp. 1–16.

[77] H. Asghari, M. J. van Eeten, and J. M. Bauer, "Economics of fighting botnets: Lessons from a decade of mitigation," *IEEE Security & Privacy*, vol. 13, no. 5, pp. 16–23, 2015.

[78] Messaging Anti-Abuse Working Group and others, "M3AAWG best practices for the use of a walled garden," 2015. [Online]. Available: https://www.m3aawg.org/documents/en/m3aawg-best-common-practices-use-walled-garden-version-20

[79] International Telecommunication Union (ITU), "ITU Botnet Mitigation Toolkit," https://www.itu.int/ITU-D/cyb/cybersecurity/projects/botnet.html, 2018.

[80] K. Krol, M. Moroz, and M. A. Sasse, "Don't work. can't work? why it's time to rethink security warnings," in *risk and security of internet and systems (CRiSIS), 2012 7th International conference on*. IEEE, 2012, pp. 1–8.

[81] H. Almuhimedi, A. P. Felt, R. W. Reeder, and S. Consolvo, "Your reputation precedes you: History, reputation, and the chrome malware warning," in *Symposium on Usable Privacy and Security (SOUPS)*, vol. 4, 2014, p. 2.

[82] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor, "Crying wolf: An empirical study of ssl warning effectiveness." in *USENIX security symposium*, 2009, pp. 399–416.

[83] A. Mathur, J. Engel, S. Sobti, V. Chang, and M. Chetty, "They Keep Coming Back Like Zombies": Improving Software Updating Interfaces," in *SOUPS*, 2016, pp. 43–58.

[84] S. Kim and M. S. Wogalter, "Habituation, dishabituation, and recovery effects in visual warnings," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 53, no. 20. Sage Publications Sage CA: Los Angeles, CA, 2009, pp. 1612–1616.

[85] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, "The emperor's new security indicators," in *Security and Privacy, 2007. SP'07. IEEE Symposium on*. IEEE, 2007, pp. 51–65.

[86] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter, "Your attention please: designing security-decision UIs to make genuine risks harder to ignore," in *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM, 2013, p. 6.

[87] C. Bravo-Lillo, L. Cranor, S. Komanduri, S. Schechter, and M. Sleeper, "Harder to ignore," *Revisiting pop-up fatigue and approaches to prevent it, USENIX Association*, pp. 105–111, 2014.

[88] R. Wash, E. Rader, K. Vaniea, and M. Rizor, "Out of the loop: How automated software updates cause unintended security consequences," in *Symposium on Usable Privacy and Security (SOUPS)*, 2014, pp. 89–104.

[89] M. Fagan and M. M. H. Khan, "Why do they do what they do?: A study of what motivates users to (not) follow computer security advice," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016, pp. 59–75.

[90] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. Telang, "Do or do not, there is no try: user engagement may not improve security outcomes," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016, pp. 97–111.

[91] P. Black, I. Gondal, and R. Layton, "A survey of similarities in banking malware behaviours," *Computers & Security*, 2017.

[92] O. Cetin, C. Ganán, M. Korczynski, and M. van Eeten, "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)*, 2017.

[93] O. Çetin, L. Altena, C. Gañán, T. Kasama, D. Inoue, K. Tamiya, Y. Tie, K. Yoshioka, and M. van Eeten, "Cleaning Up the Internet of Evil Things: Real-World Evidence on ISP and Consumer Efforts to Remove Mirai," in *The Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, 2019.

[94] O. Çetin, L. Altena, C. Gañán, and M. van Eeten, "Let Me Out! Evaluating the Effectiveness of Quarantining Compromised Users in Walled Gardens," in *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. Baltimore, MD: USENIX Association, 2018. [Online]. Available: https://www.usenix.org/conference/soups2018/presentation/cetin

[95] L. Eduardo and R. Montoro, "mDNS - Telling the world about you (and your device)," https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/mdns-telling-the-world-about-you-and-your-device/, 2012.

[96] N. Aviram, S. Schinzel, J. Somorovsky, N. Heninger, M. Dankel, J. Steube, L. Valenta, D. Adrian, J. A. Halderman, V. Dukhovni *et al.*, "DROWN: breaking TLS using SSLv2," in *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, 2016, pp. 689–706.

[97] K. Soska and N. Christin, "Automatically detecting vulnerable websites before they turn malicious," in *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, 2014, pp. 625–640.

[98] G. Pellegrino, O. Catakoglu, D. Balzarotti, and C. Rossow, "Uses and Abuses of Server-Side Requests," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2016, pp. 393–414.

[99] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai Botnet," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, 2017, pp. 1093–1110.

[100] R. Wash and E. J. Rader, "Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users," in *Eleventh Symposium On Usable Privacy and Security, SOUPS 2015, Ottawa, Canada, July 22-24, 2015.*, 2015, pp. 309–325.

[101] TeleGeography, "Telegeography Globalcomms Data," 2017. [Online]. Available: http://shop.telegeography.com/products/globalcomms-database

[102] Shadowserver Foundation, "Shadowserver Reports," 2018. [Online]. Available: https://www.shadowserver.org/wiki/pmwiki.php/Services/Reports

[103] Y. M. P. Pa, S. Suzuki, K. Yoshioka, T. Matsumoto, T. Kasama, and C. Rossow, "IoTPOT: Analysing the Rise of IoT Compromises," in *9th USENIX Workshop on Offensive Technologies (WOOT 15)*. Washington, D.C.: USENIX Association, 2015.

[104] G. F. Lyon, *Nmap network scanning: The official Nmap project guide to network discovery and security scanning*. Insecure, 2009.

[105] D. Holmes, "The Mirai Botnet Is Attacking Again...)," 2018. [Online]. Available: https://www.darkreading.com/partner-perspectives/f5/the-mirai-botnet-is-attacking-again/a/d-id/1331031

[106] J. B. Ullrich, "An Update On DVR Malware: A DVR Torture Chamber," SANS Technology Institute, Tech. Rep., 2017. [Online]. Available: https://isc.sans.edu/forums/diary/An+Update+On+DVR+Malware+A+DVR+Torture+Chamber/22762/

[107] Federal Bureau of Investigation (FBI), "Foreign cyber actors target home and office routers and networked devices worldwide," 2018, https://www.ic3.gov/media/2018/180525.aspx.

[108] B. Botezatu, "Hide and Seek IoT Botnet resurfaces with new tricks, persistence," Bitdefender Labs, 2018. [Online]. Available: https://labs.bitdefender.com/2018/05/hide-and-seek-iot-botnet-resurfaces-with-new-tricks-persistence/

[109] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The Menlo Report," *IEEE Security & Privacy*, vol. 10, no. 2, pp. 71–75, 2012.

[110] K. K. e Silva, "How industry can help us fight against botnets: notes on regulating private-sector intervention," *International Review of Law, Computers & Technology*, vol. 31, no. 1, pp. 105–130, 2017.

[111] E. Tjong Tjin Tai, B.-J. Koops, D. Op Heij, K. Silva, and I. Škorvánek, "Duties of care and diligence against cybercrime," *Tilburg Private Law Working Paper*, 2017.

[112] A. Florini, "Making transparency work," *Global Environmental Politics*, vol. 8, no. 2, pp. 14–16, 2008.

[113] W. Powell, "Neither market nor hierarchy," *The sociology of organizations: classic, contemporary, and critical readings*, vol. 315, pp. 104–117, 2003.

[114] T. Tenbensel, "Multiple modes of governance: Disentangling the alternatives to hierarchies and markets," *Public Management Review*, vol. 7, no. 2, pp. 267–288, 2005.

[115] ICANN, "ICANN-Accredited Registrars," 2019. [Online]. Available: https://www.icann.org/registrar-reports/accredited-list.html

[116] WHOIS Accuracy Program Specification, 2019, https://www.icann.org/resources/pages/approved-with-specs-2013-09-17-en#whois-accuracy.

[117] ANP XTRA, "Boete bedrijven bij lakse aanpak kinderporno," 2019, https://www.telegraaf.nl/nieuws/1500517321/boete-bedrijven-bij-lakse-aanpak-kinderporno.

[118] M. van Eeten, Q. Lone, G. Moura, H. Asghari, and M. Korczyński, "Evaluating the impact of AbuseHUB on botnet mitigation," *arXiv preprint arXiv:1612.03101*, 2016.

APPENDIX A

Content of abuse reports and cleanup Website

A.1 Example of anti-malware organization e-mail notification

Subject: Malware URL notification - poorcompromisedwebsite .com

Body:hxxp://poorcompromisedwebsite .com/user.php is currently being abused to spread malware. This means it may be placing Internet users at risk. Please investigate and take appropriate action to resolve or mitigate the threat.

Description: Asprox botnet dropper
Date/time of detection: 2014-12-07 at 00.31 (GMT+1)
IP address at time of detection: 195.158.28.146
Additional parties notified: abuse@poorcompromisedwebsite.com (site owner)

You are receiving this report because this was listed as the technical contact e-mail in the WHOIS record for 195.158.28.146. If you believe you have received this report in error, or for more information, please contact us at this address: abuse-reporter@stopbadware.org
Caution: Opening malware URLs in your browser can infect your computer. For security reasons, URLs in this e-mail have been modified by replacing http

with hxxp and by adding a space before the first dot(.)
==========
ADDITIONAL INFORMATION
==========
Detailed malware description:
URL accessed: hxxp://poorcompromisedwebsite .com/user.php?c=Rw
Behaviour: Delivers malicious executables and ZIP files.
Special condition: Only delivers malicious executables when accessed through Windows Internet Explorer.

Tips for cleaning securing a compromised website:
https://www.stopbadware.org/asprox-cleanup-advice7NSVRLZ

A.2 Example of University e-mail notification

Subject: Malware URL notification - poorcompromisedwebsite .com
Body:hxxp://poorcompromisedwebsite .com/user.php is currently being abused to spread malware. This means it may be placing Internet users at risk. Please investigate and take appropriate action to resolve or mitigate the threat.

Description: Asprox botnet dropper
Date/time of detection: 2014-12-07 at 00.31 (GMT+1)
IP address at time of detection: 10.1.5.3
Additional parties notified: abuse@poorcompromisedwebsite.com (site owner)

You are receiving this report because this was listed as the technical contact e-mail in the WHOIS record for 10.1.5.3. If you believe you have received this report in error, or for more information, please contact us at this address: malwarereporter-tbm@tudelft.nl.

Caution: Opening malware URLs in your browser can infect your computer. For security reasons, URLs in this e-mail have been modified by

replacing http with hxxp and by adding a space before the first dot(.)
==========
ADDITIONAL INFORMATION
==========
Detailed malware description:
URL accessed: hxxp://poorcompromisedwebsite .com/user.php?c=O
Behaviour: Delivers malicious executables and ZIP files.
Special condition: Only delivers malicious executables when accessed through Windows Internet Explorer.

Tips for cleaning securing a compromised website:
http://www.cleanup-advice.tudelft.nl/WJUB5TG

A.3 Example of individual researcher e-mail notification

Subject: Malware URL notification - poorcompromisedwebsite .com
Body:hxxp://poorcompromisedwebsite .com/error.php is currently being abused to spread malware. This means it may be placing Internet users at risk. Please investigate and take appropriate action to resolve or mitigate the threat.

Description: Asprox botnet dropper
Date/time of detection: 2014-12-07 at 00.31 (GMT+1)
IP address at time of detection: 112.78.8.33
Additional parties notified: abuse@poorcompromisedwebsite.com (site owner)

You are receiving this report because this was listed as the technical contact e-mail in the WHOIS record for 112.78.8.33. If you believe you have received this report in error, or for more information, please contact us at this address: malwarereporting@gmail.com.

Caution: Opening malware URLs in your browser can infect your computer. For security reasons, URLs in this e-mail have been modified by replacing http with hxxp and by adding a space before the first dot(.)
==========
ADDITIONAL INFORMATION
==========
Detailed malware description:
URL accessed: hxxp://poorcompromisedwebsite .com/error.php?c=W
Behaviour: Delivers malicious executables and ZIP files.
Special condition: Only delivers malicious executables when accessed through Windows Internet Explorer.

Tips for cleaning & securing a compromised website:
http://cleanup-advice.besaba.com/#MNVTUUT

A.4 StopBadware cleanup websites

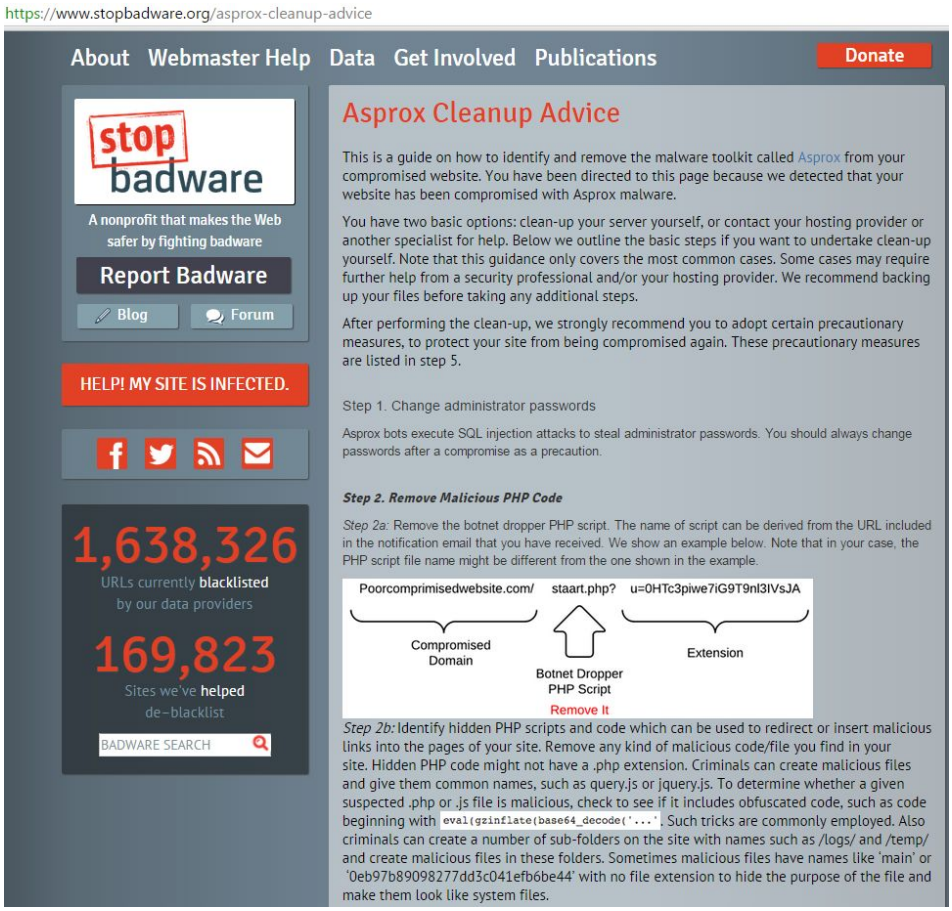


Figure A.1: Cleanup website for high reputation group

A.5 University cleanup websites

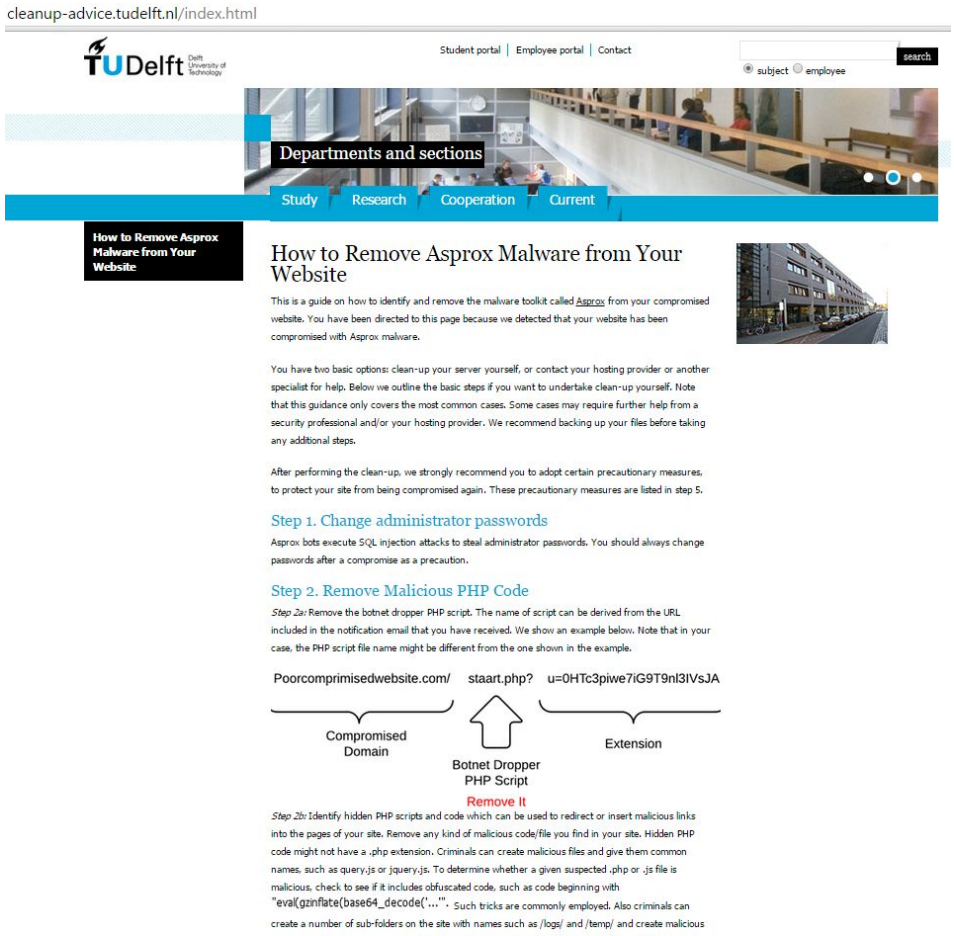


Figure A.2: Cleanup website for medium reputation group

A.6 Free hosting cleanup websites

cleanup-advice.besaba.com/How-to-Clean-Up-an-Infected-Website/

HOME | HOW TO REMOVE ASPROX MALWARE FROM YOUR WEBSITE | EXTERNAL CLEAN UP AND UPDATE SITES | CONTACTS

How to Remove Asprox Malware from Your Website

This is a guide on how to identify and remove the malware toolkit called Asprox from your compromised website. You have been directed to this page because we detected that your website has been compromised with Asprox malware.

You have two basic options: clean-up your server yourself, or contact your hosting provider or another specialist for help. Below we outline the basic steps if you want to undertake clean-up yourself. Note that this guidance only covers the most common cases. Some cases may require further help from a security professional and/or your hosting provider. We recommend backing up your files before taking any additional steps.

After performing the clean-up, we strongly recommend you to adopt certain precautionary measures, to protect your site from being compromised again. These precautionary measures are listed in step 5.

Step 1. Change administrator passwords

Asprox bots execute SQL injection attacks to steal administrator passwords. You should always change passwords after a compromise as a precaution.

Step 2. Remove Malicious PHP Code

Step 2a: Remove the botnet dropper PHP script. The name of script can be derived from the URL included in the notification email that you have received. We show an example below. Note that in your case, the PHP script file name might be different from the one shown in the example.

Poorcomprimisedwebsite.com/ staart.php? u=0HTc3piwe7iG9T9nl3IVsJA

Compromised Domain | Botnet Dropper PHP Script | Extension

Remove It

Step 2b: Identify hidden PHP scripts and code which can be used to redirect or insert malicious links into the pages of your site. Remove any kind of malicious code/file you find in your site. Hidden PHP code might not have a .php extension. Criminals can create malicious files and give them common names, such as query.js or jquery.js.

To determine whether a given suspected .php or .js file is malicious, check to see if it includes obfuscated code, such as code beginning with "eval(gzinflate(base64_decode("...". .

Such tricks are commonly employed. Also criminals can create a number of sub-folders on the site with names such as /logs/ and /temp/ and create malicious files in these folders. Sometimes malicious files have names like "main" or "0eb97b89098277dd3c041efb6be44" with no file extension to hide the purpose of the file and make them look like system files.

More examples of malicious PHP code can be seen here: http://aw-snap.info/articles/php-examples.php.

Step 3. Remove hidden HTML Elements

Placing hidden malicious links on the pages of websites is a common tactic among cyber-criminals. Hackers will place the links in a html element that can be "hidden" using CSS such as a < div> , < span> < iframe> or even a list < ul>. This type of hack is fairly easy to spot when viewing the source code of the page (type Ctrl+U in most browsers). To clean up this hack, simply delete the malicious links from the pages on the website. Commonly, these hidden elements can be found at the very beginning of the file, before the document type declaration or < html> tag and/or after the closing < /html> tag

Figure A.3: Cleanup website for low reputation group

APPENDIX B

Vulnerability notification, survey and website contents

B.1 Conventional notification content for network operators and nameserver operators

Subject: Vulnerable DNS Nameserver at ns1.example.com

Body: Cybersecurity researchers from Delft University of Technology have been conducting scans to identify DNS nameservers that are vulnerable to an attack called “zone poisoning”. The vulnerability allows an attacker to replace, add and remove Resource Records in authoritative zone files on the nameserver. In practice, this means an attacker can point the domain name to an IP address under the attacker’s control, add subdomains, or point existing subdomains, such as for email or ssh, to other IP addresses.

We scanned for this vulnerability by sending a single RFC-compliant DNS packet to all publicly visible nameservers. The response of your name server indicated that it is vulnerable to malicious dynamic updates. We did not exploit the server or interact with the existing records on it.

We have observed the following vulnerable nameservers on your network:

```
ns1.example.com
ns2.example.com
```

What can you do about this problem? The vulnerability can be mitigated by using an access control list on your name server, though this can still be circumvented via IP spoofing since the attack only needs a single UDP packet. The secure solution is to either disable so-called 'dynamic updates' or to enable Transaction Signatures (TSIG) on the server and permitting only DNS dynamic updates with authorized keys.

Did you find this notification useful? Or do you object to these kinds of scans? We are doing research to make vulnerability and abuse notifications more effective for network operators and domain owners. Please help us to make them better for everyone by taking a 5 minute anonymous survey at: `http://www.surveylink.com/[surveylink]`

You can leaves us feedback via the survey or contact us directly at `vulnerabilityreporter@tudelft.nl`.

Thank you!

TU Delft Security Notifications Project

****

List of vulnerable domains:

```
example1.com
example2.com
example3.com
```

B.2 Demonstrative notification content for network operators and nameserver operators

Subject: Vulnerable DNS Nameserver at ns1.example.com

Body: Cybersecurity researchers from Delft University of Technology have been conducting scans to identify DNS nameservers that are vulnerable to an attack called “zone poisoning”. The vulnerability allows an attacker to replace, add and remove Resource Records in authoritative zone files on the nameserver. In practice, this means an attacker can point the domain name to an IP address under the attacker’s control, add subdomains, or point existing subdomains, such as for email or ssh, to other IP addresses.

We scanned for this vulnerability by sending a single RFC-compliant DNS packet to all publicly visible nameservers. The response of your name server indicated that it is vulnerable to malicious dynamic updates. We did not exploit the server or interact with the existing records on it.

We have observed the following vulnerable nameservers on your network:

`ns1.example.com`

You can safely and easily test the vulnerability of your name server on our website at `zonepoisoning.com`. To prevent others from using the tool to search for vulnerable nameservers, we provide you with a unique token. Please use this URL to test domains using your nameserver(s):

`http://zonepoisoning.com/[uniquecode]`

You can use any of the vulnerable domain names mentioned at the bottom of this email to test the vulnerability, for example: example.com

Our website provides a simple interface that lets you add an innocent resource record to your nameserver for the subdomain ‘zonepoisoning’ – for example `zonepoisoning.example1.com`.us. If the benign subdomain is successfully added, it means your server is vulnerable and all existing records can be changed from anywhere on the Internet! You can also use our website to check whether the vulnerability has been fixed.

What can you do to fix this problem? The vulnerability can be miti-

gated by using an access control list on your name server, though this can still be circumvented via IP spoofing since the attack only needs a single UDP packet. The secure solution is to either disable so-called 'dynamic updates' or to enable Transaction Signatures (TSIG) on the server and permitting only DNS dynamic updates with authorized keys.

Did you find this notification useful? Or do you object to these kinds of scans? We are doing research to make vulnerability and abuse notifications more effective for network operators and domain owners. Please help us to make them better for everyone by taking a 5 minute anonymous survey at: `http://www.surveylink.com/[surveylink]`

You can leaves us feedback via the survey or contact us directly at vulnerabilityreporter@tudelft.nl.

Thank you!

TU Delft Security Notifications Project

****

List of vulnerable domains:

```
example1.com
example2.com
example3.com
```

B.3 Destination of injected record

Welcome to the "DNS dynamic update measurement" server project

Some questions you may have:

1. What computer is this?
   - This is a computer from the Economics of CyberSecurity group from Delft University of Technology.
   - It is run by researchers from this group.
2. Why are we sending the DNS dynamic update packets to your server?
   - We carry out Internet-wide measurements in order to track the population of servers vulnerable to the nonsecure DNS dynamic updates and notify the system operators responsible for misconfigured machines.
3. Where can I learn more about secure DNS dynamic updates?
   - Please visit this website (ISC BIND version 9.3) or this website (Windows Server 2008) for more details.
4. What update do you send to my DNS server?
   - The following A record: `researchdelft.2ndLevelDomainName 86400 A 192.42.131.1`
5. OK, if it is for research, I have no problem with it.
   - Thanks so much! We appreciate it.
6. No, I want you to stop sending the DNS dynamic update packets to my server.
   - Please just write to us (maciej [dot] korczynski [a_t] tudelft [dot] nl) (PGP: 848571D0) and we will include your zone/name server in our not-scan list.

B.4 Survey questionnaire

Security Notification Survey

Please help us improve security notifications by answering a 2-minute anonymous survey. Each question is optional, please answer the ones that you feel comfortable with. Your feedback is very important to us and we really appreciate your time.

Common Questions

1. Did your organization take prior actions to resolve the security issue before our notification?
   (a) Yes
   (b) No
2. Is your organization planning on resolving the security issue?
   (a) Yes
   (b) No
3. Do you feel it was acceptable for us to scan the nameserver for this security issue?
   (a) Yes
   (b) No
4. Do you feel it was acceptable for us to notify your organization?
   (a) Yes
   (b) No
5. Would your organization want to receive similar security vulnerability/misconfiguration notifications in the future?
   (a) Yes
   (b) No
6. Did we notify the correct contact?
   (a) Yes
   (b) No

7. Is there anything you want to tell us about our scans, notifications or any other issue related to this research or to security notifications in general?

Specific Questions to Nameserver Operators and Network Operators

1. How would you characterize your organization?
   (a) Hosting provider
   (b) Reseller
   (c) DNS server provider (Only in 1st Campaign)
   (d) ISP broadband
   (e) Content delivery/distribution network
   (f) Registrar
   (g) Other - Write In ...
2. How many employees work at your organization?
   (a) 1
   (b) 2-24
   (c) 25-99
   (d) 100-499
   (e) 500-999
   (f) 1,000+
   (g) Other - Write In ...

Specific Questions to Domain Owners

1. You are the contact for this domain. How would you characterize yourself?
   (a) I am an individual who owns this domain
   (b) I am a member of the organization who owns this domain
   (c) Other - Write In ...
2. If the domain is owned by an organization, how large is this organization?
   (a) 1-24
   (b) 25-99

(c) 100-499
(d) 500-999
(e) 1,000+
(f) Other - Write In ...

Specific Questions for Network Operators and Domain Owners

1. Is your organization in charge of maintaining name server?
   (a) Yes
   (b) No

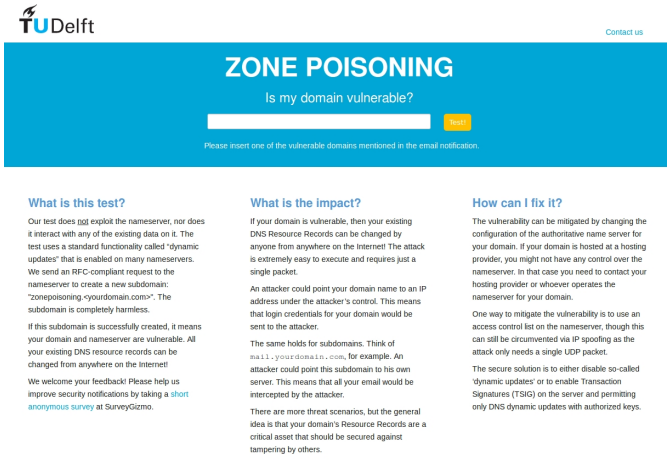
Specific Questions for the Demo Group

1. We set up the site zonepoisoning.com to enable you to safely demonstrate the security issue. Do you feel this was useful?
   (a) I went to the site, but I did not find it useful
   (b) I found it somewhat useful
   (c) I found it very useful
   (d) I did not go to the site
   (e) Other - Write In ...

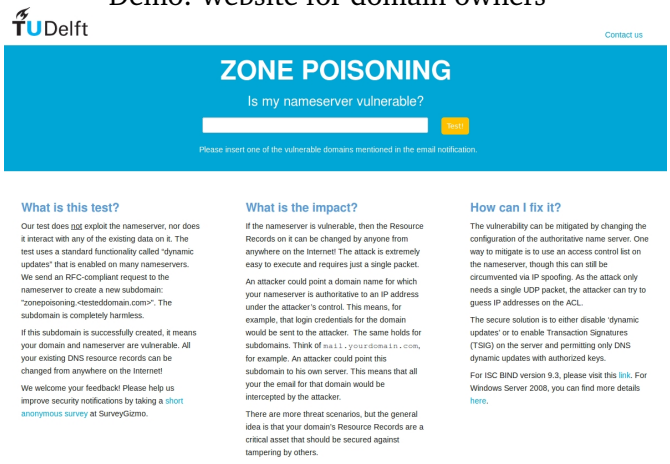
Specific Questions for the Conventional Notification Group

1. Would it have been useful if we had provided you with a site where you could safely test and demonstrate the security issue?
   (a) Not useful
   (b) Somewhat useful
   (c) Very useful
   (d) Don't know
   (e) Other - Write In ...

B.5 Vulnerability demonstration website



Demo. website for domain owners

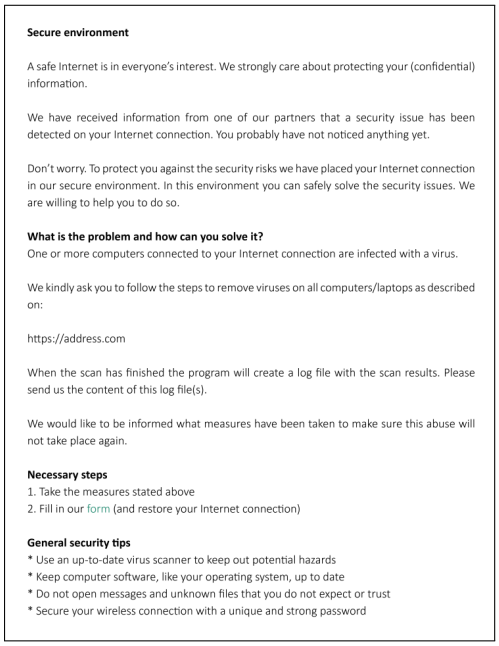


Demo website for nameserver operators

APPENDIX C

Content of walled garden notifications for malware

C.1 Walled garden landing page



Secure environment

A safe Internet is in everyone's interest. We strongly care about protecting your (confidential) information.

We have received information from one of our partners that a security issue has been detected on your Internet connection. You probably have not noticed anything yet.

Don't worry. To protect you against the security risks we have placed your Internet connection in our secure environment. In this environment you can safely solve the security issues. We are willing to help you to do so.

What is the problem and how can you solve it?
One or more computers connected to your Internet connection are infected with a virus.

We kindly ask you to follow the steps to remove viruses on all computers/laptops as described on:

https://address.com

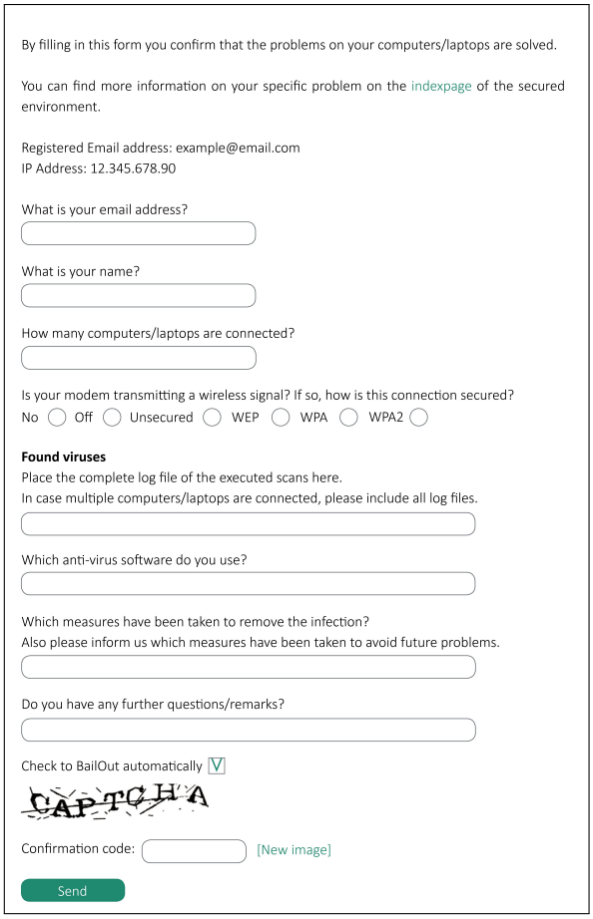
When the scan has finished the program will create a log file with the scan results. Please send us the content of this log file(s).

We would like to be informed what measures have been taken to make sure this abuse will not take place again.

Necessary steps
1. Take the measures stated above
2. Fill in our form (and restore your Internet connection)

General security tips
* Use an up-to-date virus scanner to keep out potential hazards
* Keep computer software, like your operating system, up to date
* Do not open messages and unknown files that you do not expect or trust
* Secure your wireless connection with a unique and strong password

C.2 Walled garden release form



By filling in this form you confirm that the problems on your computers/laptops are solved.

You can find more information on your specific problem on the indexpage of the secured environment.

Registered Email address: example@email.com
IP Address: 12.345.678.90

What is your email address?

What is your name?

How many computers/laptops are connected?

Is your modem transmitting a wireless signal? If so, how is this connection secured?
No ◯ Off ◯ Unsecured ◯ WEP ◯ WPA ◯ WPA2 ◯

Found viruses
Place the complete log file of the executed scans here.
In case multiple computers/laptops are connected, please include all log files.

Which anti-virus software do you use?

Which measures have been taken to remove the infection?
Also please inform us which measures have been taken to avoid future problems.

Do you have any further questions/remarks?

Check to BailOut automatically ☑

CAPTCHA

Confirmation code: [New image]

Send

APPENDIX D

Content of walled garden notifications for vulnerabilities

D.1 Open DNS resolver walled garden notification content

Secure environment

A safe Internet is in everyone's interest. We strongly care about protecting your (confidential) information.

We have received information from one of our partners that a security issue has been detected on your Internet connection. You probably have not noticed anything yet.

Don't worry. To protect you against the security risks we have placed your Internet connection in our secure environment. In this environment you can safely solve the security issues. We are willing to help you to do so.

What is the problem and how can you solve it?

Your Internet Connection is hosting a DNS Server.

This DNS Server is currently acting as a "Open Resolver". These kind of servers can be used to perform DDos Attacks. It is important that you take immediate action. One possibility is that you have installed a badly configured DNS Server yourself. Also these problems can be caused by a modem that acts like an "Open Resolver".

When you have installed the DNS server yourself please remove it from your Internet connection as soon as possible. The problem is then solved immediately.

When you are using your own modem please check and change the current configuration as soon as possible. It is import that the DNS functionality is no longer present. Please check the manual of your modem in case you need help. You can also contact your modem vendor. It is a possibility that you have to renew the firmware of your modem. Another immediate solution is reconnecting the modem that was provided by our company.

Necessary steps

1. Take the measures stated above
2. Fill in our form (and restore your Internet connection)

General security tips

*Use an up-to-date virus scanner to keep out potential hazards
*Keep computer software, like your operating system, up to date
*Do not open messages and unknown files that you do not expect or trust
*Secure your wireless connection with a unique and strong password

D.2 mDNS walled garden notification content

Secure environment

A safe Internet is in everyone's interest. We strongly care about protecting your (confidential) information.

We have received information from one of our partners that a security issue has been detected on your Internet connection. You probably have not noticed anything yet.

Don't worry. To protect you against the security risks we have placed your Internet connection in our secure environment. In this environment you can safely solve the security issues. We are willing to help you to do so.

What is the problem and how can you solve it?

At this moment your Internet connection can be used for sending a large number of malicious requests to other Internet users. These requests can form a flood of data that is capable of entirely shutting off the Internet connection of the victim. This problem is probably caused by a misconfiguration in your router. Possibly you have enabled the option DMZ (Default Server) or UPnP in your router.

If you are using ISP?s router you can solve this problem by resetting your router to factory defaults.

In case your are using a privately owned router please connect the ISP provided router instead. If you do not have the technical skills to solve this problem yourself please contact a professional like your computer vendor or IT partner.

Necessary steps

1. Take the measures stated above
2. Fill in our form (and restore your Internet connection)

General security tips

*Use an up-to-date virus scanner to keep out potential hazards
*Keep computer software, like your operating system, up to date
*Do not open messages and unknown files that you do not expect or trust
*Secure your wireless connection with a unique and strong password

APPENDIX E

Content of walled garden notifications for infected IoT devices

E.1 Standard walled garden notification content

Secure environment
A safe Internet is in everyone's interest. We strongly care about protecting your (confidential) information.
We have received information from one of our partners that a security issue has been detected on your Internet connection. You probably have not noticed anything yet.
Don't worry. To protect you against the security risks we have placed your Internet connection in our secure environment. In this environment you can safely solve the security issues. We are willing to help you to do so.
What is the problem and how can you solve it?
One or more devices connected to your Internet connections are infected with the Mirai-virus. This virus targets devices that make use of your Internet connection independently. In most cases IP Cameras or Digital TV decoders.
The infection probably occurred due to the use of a standard password username combination to access the device.
To solve this problem please reset all your devices to factory defaults. After the reset change all the passwords for accessing the devices to strong passwords.
In case the device can be reached by Telnet or SSH please also change these passwords.
Necessary steps
1. Take the measures stated above

2. Fill in our form (and restore your Internet connection)
General security tips
* Use an up-to-date virus scanner to keep out potential hazards
* Keep computer software, like your operating system, up to date
* Do not open messages and unknown files that you do not expect or trust
* Secure your wireless connection with a unique and strong password

E.2 Improved walled garden notification content

Secure environment
A safe Internet is in everyone's interest. We strongly care about protecting your (confidential) information.
We have received information from one of our partners that a security issue has been detected on your Internet connection. You probably have not noticed anything yet.
Don't worry. To protect you against the security risks we have placed your Internet connection in our secure environment. In this environment you can safely solve the security issues. We are willing to help you to do so. What is the problem and how can you solve it?
One or more Internet connected devices in your home have been infected with the Mirai virus. We cannot detect which Internet connected device has been infected. Most likely it is a digital video recorder (DVR), security camera or a printer connected to the Internet rather than a computer, laptop, tablet or mobile phone.
What should you do to remove the Mirai virus and prevent future infections? Please follow the steps below. If you cannot complete a step, please proceed to the next one.
1. Determine which devices are connected to your Internet connection. Reminder: The Mirai virus mainly infects Internet connected devices such as a DVR, security camera or printer connected to the Internet.
2. Change the password of the Internet connected devices. Choose a password that is hard to guess. If you do not know the current password, please refer to the manual.
By following these steps, you have prevented future infections.

3. Restart the Internet connected devices by turning it off and on again. Hereafter, the Mirai virus has been removed from the memory of the devices. Now that your Internet connected devices are safe, the last steps are to protect your router/ modem.
4. Reset your modem/router to the factory settings. On https://address.com it is described how you do this for an Experia Box.
5. Set the password of your modem/router. On https://address.com it is described how you do this for an Experia box.
Warning! If remote access to a certain device is absolutely necessary, manually define port forwards in your router for this device. On https://address.com it is described how you do this for an Experia Box.
Necessary steps
1. Take the measures stated above
2. Fill in our form (and restore your Internet connection)
General security tips
* Use an up-to-date virus scanner to keep out potential hazards
* Keep computer software, like your operating system, up to date
* Do not open messages and unknown files that you do not expect or trust
* Secure your wireless connection with a unique and strong password

APPENDIX F

Authorship Contribution

This thesis presents five empirical chapters that are published in peer-reviewed venues. I am the main author in all these 5 publications. That being said, these five publications are the product of collaborative work with a variety of co-authors. It is my great pleasure to summarize all their valuable contributions below.

In the study published in both WEIS 2015 and the Journal of Cybersecurity, I conducted the experiment, collected the entire experiment data and drafted the manuscript. Carlos Gañán conducted data analysis and helped with drafting the manuscript. Michel van Eeten, Tyler Moore and Mohammad Hanif Jhaveri helped with improving the overall text of the manuscript and conceptualizing the main ideas.

In WEIS 2017 study, the main ideas discussed and conceptualized by all the co-authors. Maciej Korczyński conducted global scans for the initial notification campaigns and provided additional scans data to tracked the impact of our notifications. Carlos Gañán helped to track visits to our demonstration website. Michel van Eeten designed the demonstration website. I implemented the demonstration website, conducted the experiment, data collection, and analysis. While I handled the bulk of the writing, all co-authors helped with writing.

In SOUP 2018, EUROS&P 2019 and NDSS 2019 papers, Lisette Altena collected and analyzed the data with my help. I guided her on what needed to be collected, provided scripts to parse and analyze collected data. I handled the bulk of the writing. Carlos Gañán and Michel van Eeten contributed substantially to improving the experimental design, approach, arguments, and writeup. Samaneh Tajalizadehkhoob helped with drafting the literature review section on EUROS&P paper. In NDSS 2019 paper, we received a valuable contribution from our Japanese colleagues. Ying Tie, Katsunari Yoshioka and Takahiro Kasama helped with sharing IoTPOT data and conducting a series of in-lab tests with actual vulnerable IoT devices to see reinfection rates. Additionally, Daisuke Inoue and Kazuki Tamiya

shared darknet data to help us track the infected IoT devices in our study.