Information Systems Research (CT20A7000)

Cost dimension of spam

Authors:
Adamski Michał
Bartosiewicz Sławomir
Otorowski Marcin

Supervisor:
Erja Mustonen-Ollila

Table of contents

Abstract 2

Introduction 3

Theoretical framework 4

Concept 5

Data gathering 7

Data validity 9

Conclusion 11

References 13

Abstract

Throughout the history of email the biggest threat connected to it was spam, which widespread was drastic recently causing a lot of troubles. Thus this increasing problem made lots of people to conduct researches in this area, which we are going to use as a basis for our theoretical framework, like Lambert (2003) studies of spammers' and spam' profiles or the widespread of spam presented by Pew Internet & American Life Project (2007). The consequences driven by spam for both individuals and companies are issue to concentrate on. In our research proposal we are going to focus mainly on costs associated with spam and loss that it creates. Furthermore, the data gathering and validation process is presented in case of the Polish financial company "Money expert", pinpointing the ways they should focus on while gathering data about spam and to look for possible mistakes that may occur doing it.

Introduction

Short e-mail, that was sent by Einar Stufferud on 1st May, 1978 became history. It was probably the first spam message ever sent. Spam is basically all about sending nearly identical or even the same messages to numerous recipients, who hadn't before agreed on receiving them. First bulk message sent by Einar Stufferud was actually invitation to his birthday, but today most of spam messages are sent in commercial, advertising purposes.

As e-mail became common way of communicating, significance and growth of spamming increased. First spam message in 1978 was sent only to 600 recipients, but in 1994 first large-scale spamming actions reached millions of people signed to more than 6000 newsgroups. In next 10 years number of spam sent each day increased to 30 billion, and nearly doubled from 2005 to 2006 to 55 billion per day. In February 2007 number of spam messages sent daily exceeded 90 billion. Today, according to MAAWG, about 80% of traffic from e-mails are made by abusive spam messages.

Spammers developed successful techniques to gather valid e-mails by using large databases, populated by large-scale harvesters, bots etc. They developed the techniques to spoof different accounts, hide their identity and let spam spreading in very obstructive way. This is why we can make assumption, that amount of unwanted spam e-mails will be still increasing, as well as its share in total amount of e-mails sent each day.

The surprisingly high share of spam gives clue, how harmfully it is affecting the life of typical internet user or either performance of company, that strongly relies on communication through e-mail channel. If 80% of e-mails is spam, then without any further investigation anyone can estimate, that about 80% of physical capacities and bandwidth are wasted, and in the same way about 80% of time spent on reading e-mails is wasted, not to mention what would be potential losses of this phenomena. IT specialists started to care about ways of preventing spam, and nowadays companies and individual are putting gradually more attention to this.

In our work we would like to focus on consequences of spam for both individuals and companies, paying much attention to spam industry in Poland – country, in which new communication techniques (including internet and e-mail) are relatively young. In

this proposal we would like to find out what is real scale of spam, what are the consequences and implication of this idiosyncratic phenomena.

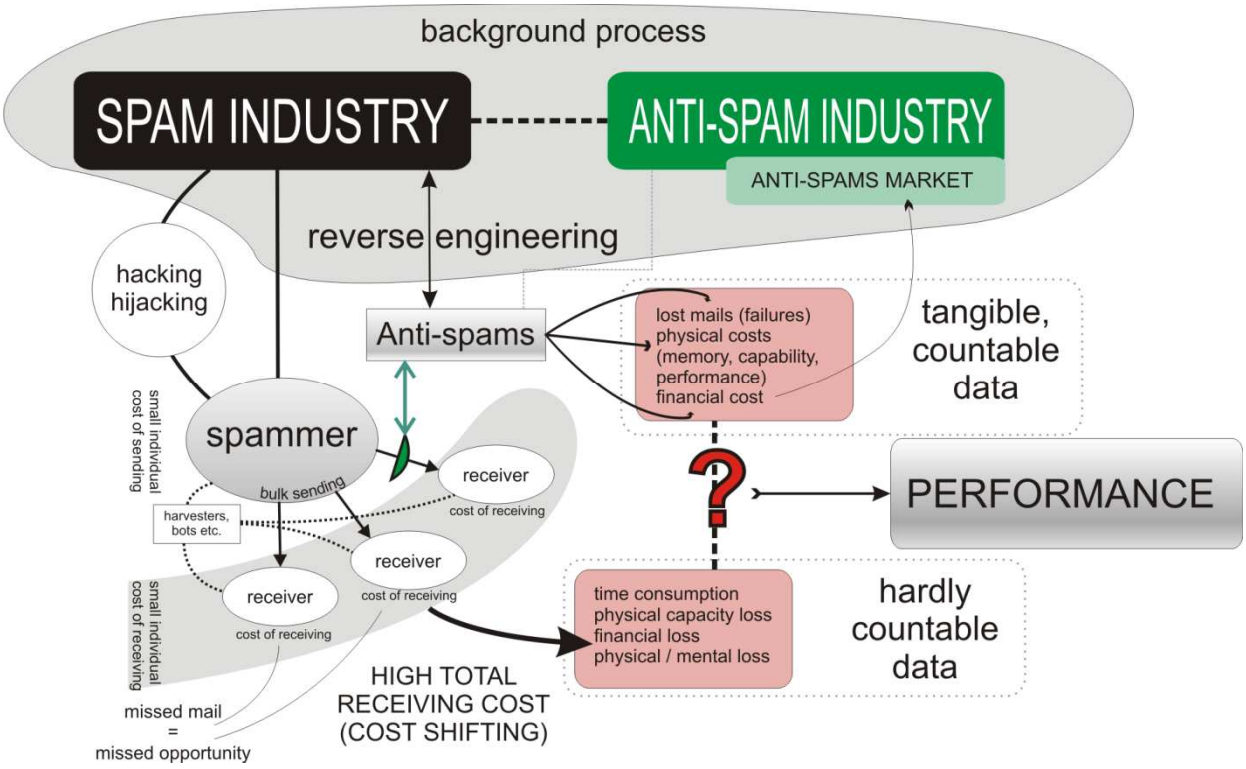
Theoretical framework

Many researches and theories about spam have already been created. In the context of this work, special attention should be focused on comprehensive research by Anselm Lambert. Lambert (2003) studied spam' and spammers' profiles, deducting number of conclusion. He claimed, that there is no universal way to stop spam, solution to this problem is based on wide approach of legislation, education and technical abilities and technologies implemented for spam filtering purposes. In context of our work special attention should be put on phenomena, which Lambert described as it follows: "Spammers continuously mutate their messages in order to defeat current spam filtering technologies". Significance of this fact will be discussed in our conceptual framework. Lambert provides the knowledge about ways of spreading the spam, helping imagine its scale and future problems.

Tremendous growth of spam and reaction of users are presented in Pew Internet & American Life Project (2007). According to this research both in personal and commercials uses, e-mail users notice more spam than ever. For personal use, 37% noticed more spam, half of them didn't see any difference, 10% noticed less spam. It is even worse for work account, where 29% of users are getting more spam, and 55% didn't see any difference. The authors of report concludes, that spam still continues to plague the internet according to data collected, in a meanwhile however despite of increased volume of spam, users are either less bothered or less aware of its influence. Furthermore, according to Dana Gardner – analyst at Yankee Group from Boston – "for enterprises, educational institutions and government, spam is a cost and liability".

Paul Judge (2003) says, that "The scale and effect suggests that spam is a type of information security problem. It has many properties in common with denial-of-service and network intrusion". His main conclusion is, that "spam is an unauthorized use of resources: bandwidth, storage, processing and people's time". This phenomena will be considered further as effect of cost switching – from spammer', to receiver's side.

Concept



This graph shows basis of our concept. The very issue, that let spam phenomena to happen is distinction between low cost of sending individual message for spammer, and its benefit for him and large overall costs for the whole web of users, which are receiving unwilling messages. Considering low cost of bulk sending, it is critical to point out, that due to large number mails sent each day (several million messages), even high costs of supporting services (bandwidth, physical facilities etc.) cost of sending one message is really insignificant. In fact, the process of switching costs from sender to receivers is the one, that allows spammer to perform. Unsurprisingly, spammer does not have to use his own computer and own bandwidth to send bulk spam – instead, he can use sophisticated methods of hacking into other victims, taking control over their machines, and use them against them by sending spam, without victim's knowledge and attention. Spam is being sent to victims, wasting bandwidth and physical capacities of hardware in duality – on server-side (receiving, storing, forwarding) and on client-side (receiving, processing, storing).

Spam can have influence performance and cause loss in number of ways. Consider for example time consumption – a time needed to browse through mails, waiting for

response of computer (larger, because larger amount of bulk spam are being processed). Another thing – physical capacity loss. Each mail stored on hard disk requires small additional capacity. Given large number of spam, amount of wasted space is tremendous. Furthermore, it implies financial loss, because each unit of time, each unit of megabyte on hard disk is cost, nowadays. All the factors connected also affect psychical intangibles, increase frustration, decrease efficiency and so on. What is important here is that all these costs are mostly intangible, so measuring them is not really simple and needs some ideas how to do this. For instance, frustration and efficiency cannot be easily measured, and no quantitative data can be received (it's risky to claim, that some decrease in efficiency is only due to spam, and it's really hard if not possible to separate spam influence and other factors to have unbiased results). The situation is a little better in financial and physical loss, as we can try to measure what is the price of factors, which are loss due to spam (price for each megabyte, price for each hour of browsing and reading spam). Nevertheless, the effect of spam are not easily measurable – we will try to find out some approach, which helps in this context.

Going back to our model, let's introduce spam filter (in this moment it is not important if it is client or server side). Spam filter is basically a software – it is represented in our graphical model between spammer and receiver. While this solution allow to decrease influence of spam described previously, we need to consider side effects of this way. First of all, it is only software, so we have to take into account possible failures of system (wrong classification, removing or hiding important and valuable messages are examples of failures). Each failure creates additional cost, this time connected to spam removing process. Besides failures, spam removing still consumes some memory and physical capacity of computer, thus it can generate financial costs (purchasing, implementing and learning software, cost of memory and capacity used, cost of failures and so on). What is different here from previous costs is, that these are easily measurable. Simple test with spam filter let check, what is it's efficiency, therefore providing further solution to this thing. Capacity and memory usage is well-known, usually provided by producer, or it can always be manually checked by mid-advance computer user.

This is encounter we want to focus in our proposal. On the one hand we have costs of receiving spam, on the other hand there costs of preventing spam. This

comparison and its possible further influence will be covered by our proposal, as one part of economic dimension of spam. A supporting case of costs in idiosyncratic company should be use here.

This would be enough, unless considering another process, which takes place simultaneously. In our model, it is presented by background processes. Spammers and anti-spam industry are not in stable environment. There are continuous desires of each side to beat another. For instance, spammers want to beat anti-spam methods in order to achieve bigger target. If they do so, anti-spam industry will response with changes to software in order to not only protect from existing techniques, but also preventing from further possible ways. The things are more complicated, because it has been said that spammers are always one step ahead of anti-spammers. Each side is using reverse engineering to discover rules and potential pitfalls of other techniques, fighting for its benefits. So, the conclusion for our model is, that once the whole system works without failures, it does not necessarily mean, it will work correctly forever. Spam war is factor, which significantly influence each individual user, both from spam and anti-spam techniques.

Data gathering

The firm we would like to focus on in our case is a Polish firm “Money expert”, which helps in financial decisions. It deals with lots of customers on daily basis, and one of the most and at the same time common way of contacting with customers is via e-mail. Economical problem the firm deals with is tremendous, as if you can imagine, the time spent on reading unnecessary e-mails by the employees is creating loss not only in money but also in efficiency they could have while not receiving bulk, unwanted, mails. We would like to focus our research on the quantity of spam they receive, how much it influences internal structure of the firm, by what we mean employees who have to deal with it all the time, and last aspect which is the way of IT involvement that deals with spam.

The data collection aspect plays crucial role in every research, because of this you are then able to analyze the problem from the scratch. In our proposal we are going to use both qualitative and quantitative research methods.

One of the methods available for us to measure the scale of spam in the company is to create so-called honeypots. These are purposefully set accounts that are used as a bait. As the employees operate not only with private customers e-mails, they are also using loads of newsletters so they could be still up-to-date in this fast, changing, environment. USENET newsgroups are also popular among them, as they use sort of investors- or financial-oriented groups, due to their interests. In this case we are going to create several accounts with different portfolios. According to Anselm Lambert (“Analysis of spam”;2003) we should create accounts dividing it by gender. It means that we should use obvious male and female names to check whether spammers find any distinction in this and if they are going to target particular sex group. Both male and female mails will be used with every newsgroup, portals, USENET groups. For instance one male and one female e-mail will be registered at USENET newsgroups and posted at least one message there, so the spam bots, harvesters, spammers could get that mail. The same method will go for newsletters, but as these accounts will be registered on the portals we will not select the box which says about obtaining other information, which could potentially lead to receive more spam. Everything is made in order not to receive spam, but as it turns out unfortunately it does not work as it intended. In this data gathering method we are going to obtain data telling us exactly from which idiosyncratic source the mails come from. If spammers for example care about sex distinction on e.g. newsgroups. We will be then able to check if the portals forewarn the rules users set by checking boxes while registration.

Not only spam itself is creating damage to our company. We should keep In mind also our employees who deal with it every single day and pay attention to more intangible data. Internal dissatisfaction made by spam may be very crucial from the point of employer view. The applicable method which could help to determine this data might be a survey. In this questionnaire labor will be asked about inconvenience created by the spam, how much time it takes them to go through this litter and separate it from the real customer mails, how they try to deal with it and avoid getting spam and so forth. This type of data is hard to measure, but it should also be taken into consideration, inasmuch it is employees who are facing the problem and ignorance of this topic would certainly lead to labor discontent, but as mentioned before it influences the efficiency, and the time they could devote to other customers.

Nonetheless, we should take into account also more tangible, hardware side of it. There are a lot of things influencing here this particular problem. The company has its own dedicated servers for e-mails, which are being monitored by external company. That is why we should contact this company in order to request all necessary data, which I am going to mention in a moment, from the archives and statistics they are obligatory to keep. The capacity required to keep all the e-mails on the server comes here as the first, because of a simple fact that we should analyze how much space is needed for this liter on the hard drive. Thus, the firm should provide us daily statistics about incoming mails, how much spam was classified at the beginning and how much was added due to employee intervention. This statistics will help us to analyze the trends about incoming spam and let us know how much exactly space is needed in our storage system. The next point in our data collection from server would be gathering all, daily and from last year as well, reports according to the bandwidth usage by mail-side of server. This reports will pinpoint us exactly how the traffic on our internet connection is occupied by mails. After gathering this data, comparing it with capacity used by mails, we would be then able to assess the percentage usage of bandwidth that downloaded spam mails generate. To sum up, keeping unwanted mails makes the storage costs bigger, because if we did not have any spam, we would be able to keep more e-mails from our customers. The same situation goes for internet connections, maybe if we had not received so much unnecessary mails we could have lower bandwidth? All in all, every of this aspects creates unwanted loss of costs, which may be simple spent on something else. Having in mind this obstacles, we need to collect all the data connected to this and analyze it in order to get to know what economical loss it creates.

Data validity

After gathering all the required data, we wanted to measure, we have to validate it, which means to find potential places where a mistake could be made. The most common place a mistake could occur is the survey, as the data collected there is more qualitative, questions asked there are about feelings. That is why, while preparing our survey we have to remember to make the questions clear for the reader, so that any miss interpretations would not happen, as this would lead to gathering wrong data which would conclude in false results. The other thing is the

length of the survey, because as some people like to fill surveys some people feel irritated, because they think that it is waste of their time, that is why our survey should not take too much time, approximately 10 minutes. The other thing is that the layout should be clearly readable and nice to look at so that the questionnaire would not make people annoyed because of the layout or because of wrongly stated questions. All of these efforts are to make the questionnaire more pleasant to fill out.

The questions supposed to be stated in such a way that it would be easy to compare the results, no open questions, as this would led to heterogeneity of collected data and that would mean impossible to compare them. Questions like 'Does spam annoy you?' give your answer in a scale from 1 to 10, are the kind of questions suitable for the research as they are easy to count and validate.

There are many things that has to be made in order to build a good survey, that is why we need long preparations or a good team that would create the survey for us. Building a good survey is one of critical objects of data gathering and it is most probable to make a mistake that is why we should put a lot of effort in preparing it. None the less the questionnaires' are not the most important data source in the research thy provide us witch valuable information which could help us to understand the phenomena of spam from the user point of view.

While analyzing the data we also have to remember that when people are answering questionnaires' they can make different mistakes, that is why we should for example, when analyzing the data add to the final result +/- 5%, because of the divergence of our result with actual one.

As was stated before the other tool we will use are clearly tangible data, like memory usage or bandwidth. These are audibly countable raw data which can be easily analyzed and entered into a statistical program in order to receive final data and graphs which could be then better understood. The results would concern answers to questions stated in data gathering part of the proposal.

In our research we would like also to measure the effectiveness of the partner of our target company, as the efficiency of the "Money Expert" company is pretty much dependant on that company because of outsourcing. In our CASE the company is using servers from the other company. That is why it is also important to validate the

data gathered from the other company, the problem is that we do not have any influence on the data we receive, that means we only receive raw data from the server company and we simply interpret them, but on the other hand we do not know if the data is correct and this is the place the problem is. If the data from the other company consists of any kind of mistakes it will influence our result and conduct a false report, even if we will see some deficiencies in our data analysis we will not be able to validate the data because they came from outside. That is why it is very important to receive correct data in the first place.

But as was said before both of this methods of data gathering are quantitative data which is also very tangible so actually no mistakes can occur that is why data validity is not so much concern in this area, controversially surveys, which are more quantitative data. This situation is in great favour for our research as the whole project will be based more on qualitative data which is more accurate and will provide us with better results what will conclude in better quality of the whole work.

Conclusion

Spam is the disease of the 21 century; with every year the number of spam is getting greater or even doubles. But as end-receivers begin to be more passive for this phenomenon of receiving unwanted messages the correlation between the spam and anti-spam environment is getting tightened. It is not longer end-receivers who fight against spam but it is more the anti-spam companies that are fighting with spammers.

As to understand this phenomenon we have also tried to go inside a company which is dependent on IT and internet. Our volition is to try to measure how its effectiveness is influenced by receiving spam, what kind of methods it uses to fight it and if it is succeeding. Results of our research would help the company to understand the connection of spam with costs and physical losses caused by spam which hasn't been realized before and which could on the other hand result in is saving time and money.

As our concept is very complex and lot of data has to be gathered, results of this research will reveal many unknown answers concerning connections between all the

participants of the spam environment or even show the future of spam itself and where it is going.

References

Yardena Arar, "Spam Explodes, but You Can Fight Back", published in PC World, vol. 03/2007

Helen Bowie, "The Spam Act: what does it mean?", nzbusjness.co.nz, vol 05/07

Joshua Goodman, Gordon V. Cormack, David Heckerman, "Spam and the Ongoing Battle for the Inbox", Communications of the ACM, February 2007/Vol. 50, No. 2

Edward Hurley, "Researcher chews fat on spam and security", retrieved from http://searchsecurity.techtarget.com/qna/0,289202,sid14_gci891030,00.html

Edward Hurley, "Security pros can leverage spam-busting with management", retrieved from http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci934288,00.html

Anselm Lambert, "Analysis of Spam", published in 2003 by Department of Computer Science, University of Dublin, Trinity College

Józef Muszyński, "Wolumen spamu znowu rośnie", retrieved from http://www.networld.pl/news/102274.html

Margie Semilof, "Spam - a costly nuisance", retrieved from http://searchwinit.techtarget.com/originalContent/0,289142,sid1_gci871979,00.html

Sam Vaknin, "The Economics of Spam", retrieved from http://globalpolitician.com/articledes.asp?ID=870&cid=1&sid=19