Agris on-line Papers in Economics and Informatics

Volume V Number 1, 2013

Spam as a Problem for Small Agriculture Business

A.Vasilenko, V. Očenášek

Faculty of Economics and Management, Czech University of Life Sciences in Prague, Czech Republic

Anotace

Příspěvek se zaměřuje na problematiku elektronické komunikace mezi firmou a zákazníkem. V této komunikaci je kromě jiných nástrojů využívána elektronická pošta. Problémem elektronické pošty je její snadné zneužití třetí stranou. Proto je nutné zabývat se možnostmi obrany proti nevyžádaným elektronickým zprávám. Ty s sebou nenesou pouze časovou ztrátu, ale také nebezpečí malware a phishingových útoků.

Uvedená problematika je velmi důležitá i z pohledu podnikatelských subjektů a jednotlivců také v regionech, kde se význam ICT přes existenci digitální propasti stále průběžně zvyšuje.

Klíčová slova

Email, spam, blacklist, fitrování zpráv, analýza emailových zpráv.

Abstract

The article deals with the problematic aspect of electronic communication between businesses and customesr. In this kind of communication the main tool which is used is email – electronic mail. The main problem connected with electronic email is the possibility of misuse by a third person. Because of this it’s very important to understand ways to prevent email abuse. Spam is not only about time losses, but also about risks of malware infection or phishing messages.

This topic is very important for small businesses and single agriculture subjects in rural areas. The influence of information technologies rises every year, despite the digital divide between rural areas and large cities/ industrial areas.

Key words

Email, spam, blacklist, message filtering, email analysis.

Introduction

According to Mail Anti-Abuse Working Group, about 80% of all email traveling through the internet is spam. This is a very large number, and it corresponds with end user daily problems – the mailbox contains tens of spam messages but only units of ham – real emails from real persons, customers and friends.

Basics term explanations:

- Ham – wanted emails from real persons
- Spam – unsolicited emails with „very profitable“ offers, hoax messages, malware and similar unwanted things
- Spamvertized web site or product – site or product in spam
- False positive – wanted email marked as spam – very undesirable, because the potential customer can be lost
- False negative – spam which wasn’t marked as spam, it is a problem, but not a big deal, this message must be deleted manually (José, 2012)

This is a big issue for small business, when email service is the main communication tool. Clear email boxes without spam are important for agriculture businesses. Farm markets are very popular in big cities and farms can order food through the internet. They have websites with email and customers can communicate with them (Vaněk, 2008). When these email boxes are full of spam, communication is very difficult and there is a big chance of overlooking important emails (Šimek, 2008).

Currently we have many tools that can be used as an appropriate counter measure. But the question is, are they sufficient? All of these tools stand alone. Of course, we can buy or create software which uses many of them together, but there we must be very careful, because only a blind combination might not be correct and can turn possible customers to competitor bussineses.

Material and methods

Research in this article was done at four domains which are in common usage on the internet. Within the domains all antispam software was shutdown. Every unsolicited email was moved to a special folder by the user. All messages in this folder were copied out by cron script at 01:00AM.

- Vasilenko.cz
- Jablickov.cz
- Malestranky.cz
- Nespamu.cz

There is one domain name with a similar purpose as a web site preseting agriculture business. This domain is jablickov.cz. The key task is to propagate main benefit offers to other people. Jablickov.cz offers courses for kids or mothers with kids. Agriculture businesses offer products from ecological farming (Vaněk, 2010). Spam is a special form of internet threat. There is no difference between agriculture businesses and other economical subjects.

Email header analysis

Email header contains relevant information about a message. All rows are specified in RFC for SMTP procotol. The following are important for spam analysis:

- received – there is an IP address of the sender's computer and date and time when the message was received
- subject
- body

Other rows aren‘t as important. Group analysis doesn‘t contain relevant data. Yes, there is a possibility to make a thorough analysis, but in this case, we can ignore them.

Content analysis

Spammers are sending millions and millions of emails with similar content (Alexander, 2009), (Xinyuan, 2009), (Compuoter Fraud & Securityi, 2011). When all of those messages will be the same, it would be easy to detect and delete them – ideal tools are for example the md5 hash function. Hash is an imprint of text string, when only one character in a long document has changed, the hash print is different – the principle of an electronical signature.

To defeat possible filtering based on hash, spammers put some random texts in messages. For example:

Dear **449e3d6**,:

*0.67$--Vigara

*1.71$--Levtira

*1.51$--Cilais

*1.56$--Female-Vigara

*2.12$--Family-Pack

*3.25$--Professional-Pack

http://NaK.medicclot.ru/ (random string parts are highlighted by author)

Thank you!

(Author original research, 2012)

This is one of 4972 similar emails captured between February 2011 and December 2012. All messages are different. Random parts are in the first row and at the link address. In those 4972 emails links are overall 524 domains with unique third level domain. Average cost of russian tld domain .ru is 7USD per year (Author original research, 2012), so the cost of all domains per year is 3668USD. Hosting for this domain cannot be detected, because according to who is IP tools are server in Germany or Antarctic – base MacMunro.

If we calculate an average hosting, for example VPS (virtual private server), we can assume that the cost can be about 500USD per year – the sum of costs for this spamvertized site is about 4000USD per year – based on the available date there is a relevant possibility that many more domains are registred from spamvertizing. This is the weak point of the spam rate between cost and income. All links are pointed to the same website which offers pills marked as a Canadian pharmacy. The prices and some additional texts are also different. This site is placed on 4 servers. When the IP address from this server is placed on the web browser the output is only a text string - „abab“. This spam infrastructure is hosted by cb3rob.net – known for example in the spamhaus.com project as one of top 10 spamming subjects (Alexander, 2009).

Characteristics of this spam set

As is show in this message, there are mispelled name of well know drugs:

- Vigara x Viagra
- Levtira x Levitra
- Cialis x Cilais

(Author‘s original research, 2012)

These mispelled words aren‘t mistakes. This is the countermeasure against bayesian filtering. Because bayesian filtering divides emails into single words and analyzes each word based on the share of this word in ham and spam. Common tactics include also putting some „good“ words into a spam email. All of those countermeasures are named as Bayesian poisoning (Xinyuan, 2009). Spammers are sending spam in sets – certain quantity of the same messages.

One of many similar texts:

Dear 1fad723,

** Vigara - $0.62

Levtira - $1.63

** Cilais - $1.35

Famyli Pakc - 1.90$

** Femela Vigara - 1.35

Professional Pack - 2.89$

Follow this link: http://csGkR.medicappea.ru/

Thank you, 1fad723!

In this group of messages the linked domain was the same – medicappea.ru. What is different is the 3rd domain – the only randomly generated string (Alexander, 2009).

- http://mqYz.medicappea.ru/
- http://gJsGzlj.medicappea.ru/

Other groups from this advertised web are similar – prices are randomly generated in a predefined interval. For example for Vigara (purposely mispelled Viagra) it is between 0.52 and 0.87USD (on the pages the prices are higher – for example Viagra 0.88USD).

Text patterns for these groups are mainly the same – mispelled names of pills and prices with randomly added string:

** Vigara - $0.62

++ Vigara – 0.70$

All of this is countermeasure against bayesian filtering – spammers try to make as many changes with minimal hardware consumption. Spamming is about sending a great amount of unsolicited emails with the hardware demand as low as possible.

Very dangerous for an unexperienced user are spams targeting the technology aspects of internet communication and maintaining websites. For example this message:

==

Last Call For Domain jablickov.com:

We will be offering jablickov.com for sale today. We see that you previously respond to an email about this domain, but did not submit an offer. This is your last chance to submit an offer on excelfunction.com, or we will make other arrangements.

To submit an offer of at least $97 now, click http://OCCUPYCINEMA.COM/7b82fb7d4e414868.34

But I don‘t know how much to offer!

Often people do not submit offers, because they don‘t know how much to offer. Our minimum offer price is $97. If you submit an offer of at least $97, then you will reserve your position for this domain. In almost all cases, this is enough to win the domain.

To submit an offer of at least $97 now, click here

How do I know that this is a safe transaction?

This is **a ONE-TIME** payment, after which the domain becomes your exclusive property. You never have to pay us anything for the domain ever again.

I don‘t want to rebrand everything with the new domain name

You do not have to rebrand at all. Our service includes FREE domain and email forwarding! You simply redirect the traffic from jablickov.com to doozerbrewingco.org and gain the benefit of having the preferred excelfunction.com without having to change hosting or rebranding at all.

How will I know that I own the domain?

To summarize - you can bid as low as $97, you do not pay until you receive delivery and you never have to reveal your personal payment details to anyone. This your best possible opportunity to get the preferred excelfunction.com to complement your doozerbrewingco.org domain.

Act now and get a free SEO analysis of your website (a $250 value!).

If you would rather not receive notice of these business proposals again, please click the following link, and your address will be removed immediately -http://OCCUPYCINEMA.COM/1/7b82fb7d4e414868.34

It is possible to store the mind with a million facts and still be entirely uneducated.

We are kept keen on the grindstone of pain and necessity.

== (this is shortened version of spam message – cut is made by author)

This message is based on an attempt to make the user fearful about his domain with international suffix .com. When users have no experience in IT they can be easily convinced that it is necessary to pay.

Bayesian filtering

Common tool for email analysis is Bayesian filtering, this tool is used to determine a score for each email. When the score reaches a preset level, the email is marked as „suspicious spam“ or spam. For suspicious emails the user reaction is mandatory – the user alone decides about this message. When an email is marked as spam with high probability, than the message is dropped to trash.

The key term is spamicity – the probability that this word or email is a spam. Spamicity is a number from the interval between zero and one. There are many different ways how to calculate it. But we can detect spam by patterns based on the content of spam messages. Bayesian filtering analyzes spamicity of words or small parts of a text. So this can be manipulated by adding positive words to a spam text.

Bayesian filtering for this case is not the best tool. Sets of spam messages are different and only one part of the message is similar – the link. But bayesian filtering checks only the text in the link – and the domain is very variable, so this tool is not as good as it should be. Next problem is the hardware cost. Bayesian filtering needs some cpu capacity to analyze emails and compute the final score for a message. There is some research about pre-classification spam messages to relieve some load. One future possibity can be packet analysis on middle communication node.(Muhammad, 2009)

Another way can be established by using collaborative antispam leaning system, where the cpu load is divided by the number of collaboration MTA servers. All users of this antispam network participate to make the most successful antispam collection of rules. But again – it is only about making rules and every email was analyzed as a single one.(Gu-Hsin, 2009)

Blacklisting

The effectivity of IP blacklist is low in this case. About 12% of spam messages can be filtered by this tool, but also 26 hams were blocked by an IP filter. This is the result of botnets (Alexander, 2009).When one zombie computer is in a large local net behind NAT (all computers on network communicate through one IP), all computers from this local net are affected – they are sending email from same IP, which is blocked. When big botnets have hundreds of thousands of computers under control, blacklisting is no longer an effective and reliable tool. It can be used only as auxiliary metrics.

The same situation occurs in the case of domain blocking by DNSBL. When a spam message is send, the header contains a false sender address. Therefore it cannot be considered as a reliable tool. Spam messages contain 4218 spams with the domain name jablickov.cz or vasilenko.cz. If DNSBL is applied, users from those domains cannot simply communicate between each other.

Opportunity

All antispam tools act as a single instrument. Blacklists evaluate the IP address or domain, bayesian filtering calculates the score for the entire message, DKIM or other authentification tools check the sender‘s identity. Commercial antispam solutions try to make a group out of these tools. What if there can be a compact solution to recognize spam based on identify the message as part of a single spam set? It can be easier to decide – this message is similar to several groups of spam.(Zhenhai, 2011)

Results and discussion

After 20 months of monitoring four domains, 71 572 emails were received at all of four domains in research. 95,239% of emails were spam and only 3 407 ham. This is a huge number of messages if we need to analyze and sort them manually.

When an antispam solution is applied, many of the spam messages will be at least marked as spam. But there is a big issue - can those antispam tools be trusted? How many false positive and false negative results are there and how difficult is it to set those tools to operate? And of course, can those tools help with electronic communication?

For businesses is very important to read every message from a potentional customer. So even when 68,88% (46 954) of spam are randomly generated email addresses pointing at our domain, we cannot simply say that all messeges are junk. (Author‘s original research, 2012)

In the domain the trash bin contained 98 real emails from live persons after 20 months. So it is a small amount – 0,209%, but even such small amount cannot be forgotten. (Author‘s original research, 2012)

A sample of 5000 recognized spam messages was translated with google translator to the Czech language. After translation, all messages were put again into bayesian filtering. 617 messages weren‘t recognized as spam. The big effectivity of Bayesian filtering for English written spam in Czech (or other local) language environment can be seen here. (Author‘s original research, 2012).

Advanced spam scoring

Bayesian filtering is a great tool when we received spam in a different language, fortunately for the Czech language environment. For small businesses in the Czech Republic it is one of many great tools.

More efficient filtering is using content history analysis. When we can decide based on history that a group of messages is like another, we can then efficiently defend our mailboxes against spam.

Amoeba effect and unsolicited email vector

When a spammer takes orders to spamvertize a specific product or website he gets a text with a proposal. This is the core text for a spam message. This core must be protected from bayesian analysis and the blacklist of spam words. So, several variants of this message must be ready. Changes are based on adding random strings, changing prices and adding words with a predicted positive bayesian score. This we can call an Amoeba effect – the core of this small protozoan inucellular organism is still the same, but the shape is different. For a human being it is simple to be recognized, but for a computer it is very hard.

The Amoeba effect can be mathematicaly described as a multidimensional vector. Each characteristic of a spam email have their own vector variable, lets call it UEV (Unsolicited Email Vector). The final score of the message is a composit of the sums of all vector values. Composition od UEF is based on spam characteristics – IP map address score v1, bayesian score v2, clean subject score v3 (clean – with the random string removed), link analysis v4, time characteristics v5, amount of near-like messages v6. So UEV in this simple form can be described by the equation, Where variable x is mark for single email analyzed by UEV:

$$UEV_x = v_{1x} + v_{2x} + v_{3x} + v_{4x} + v_{5x} + v_{6x}$$

For a decision to which group of spam sets belongs a single message, we must compute the vector of this message. Database for this solution can be established by the multidimensional OLAP database. (Author’s original research, 2012) (Tyrychtr, 2012).

V1

V1 vector is based on a map of IP addresses misused for sending unsolicited emails. When botnet is used for spread spam messages once, it is not at last time and there is a possibility to capture a large amount of spam messages.

V2

Differencies between sets of spam messages must be reduced by near-like message detection. This NLMD procedure eliminates some artifically added strings and signs. NLMD can be likened to database normalization - Boyce-Codd Normal Form. After NLMD body of analysed emails is clear from disturbing strings and characters, such as „*, +, /, …“, multiple spaces and aditional rows. Now v2 can be processed by bayesian filtering.

V3

The subject is an inseparable part of an email. For confusing antispam tools, spammers add random parts to the subject. The captured spam stated for example this:

- New discount <1>
- New discount <2>
- New discount <3>
- …
- New discount <10>

So for vector v3 the subjects from captured spam messages were compared with subjects in the actual message and v3 distance is calculated.

V4

The link or email for vector v4 in the text is the only way to make order of a spamvertized product, so it is a significant pointer. If we compare the link in the email with saved links from spam messages, we can decide if the link is clear or else the link is pointing to spamvertized sites. We only need to use methods for comparing web sites. Spammers cannot make infinite number of web pages – it costs money and time. The key attributes in this are the IP addresses, domains and comparing web sites with known spamvertizing sites in the database.

V5

The time and date in vector v5 can be very useful in special cases. When, for example, we have the date of January 2nd 2013 and in the email the date is January 5th 2014 or December 5th 2002, the spam is identified. Also when many near-like messages are captured with almost same date – it is very probable that it becomes one set of messages. When only one is spam, it is highly probable that all other messages are spam too.

V6

Last vector shows how many similar emails we have received. Often it is usual that we have the same messages in the mailbox for one user – it is caused by a mistake or a technical anomaly. This is not the reason to say that 3 messages with same text are spam. But 15 messages have more chance to be spam. So quantity is a last vector – it must be evaluated relatively with all 5 remaining vectors.

Conclusion

As in different computer security topics the antispam tools are one or more steps behind the spammers. At this time we only defend our mailboxes. Filtration and blocking is like pills against a headache. They cure symptoms not the cause. And we cannot cure the cause because of freedom of the internet. And we cannot restrict free access to the internet in accordance with Network neutrality. So spammers can hide behind botnets.

The only reasonable solution is based on the user. If you nobly click on a link in spam or make an order, the spam died alone. But when little fiction of internet population spends money through the spam messages, spam will be with us. As was written in this article, the main oportunity is in near-like detection. Spammers are sending very large quantity of spam, but only slightly modified by small random strings. There can be strong methods to drop them out.

When we look at received spam messages in this project, we can say that a large number is from a few sources sent by few orders. Almost the same texts, same websites at different domains. This is a way to really make spammers work hard to defeat this. When spam must have more modifications to not be recognized as similar, there is a big need of resources – generating every message as a single text is very resource consuming. Large work i salso about offensive solution to prevent spammer to send milions of emails. But this is another story.

Disadvantage of this proposal is need for large amount of spam messages. For this system is very important to build database with as much spam as possible. Second issue is need for computation power. When mail server must serve to 100 messages per hour, it can handle more deeply analysis then if have 100 messages per minute. For this, next research will be focused to benchmark UEV in real condition.

Corresponding author:
Ing. Alexandr Vasilenko
Department of Information Technologies, Faculty of Economics and Management,
Czech University of Life Sciences in Prague, Kamýcká 129, 165 21 Prague 6- Suchdol, Czech Republic
E-mail: vasilenko@pef.czu.cz

References

[1] José R. Méndez, M. Reboiro-Jato, Fernando Díaz, Eduardo Díaz, Florentino Fdez-Riverola, Grindstone4Spam: An optimization toolkit for boosting e-mail classification, Journal of Systems and Software, Volume 85, Issue 12, December 2012, Pages 2909-2920, ISSN 0164-1212, 10.1016/j.jss.2012.06.027.

[2] Vaněk, J., Jarolímek, J., Šimek, P. Development of communication infrastructure in rural areas of the Czech Republic. Agricultural Economics (Zemědělská ekonomika), 2008, year. 54, No. 3, p. 129-134. ISSN: 0139-570X.

[3] Šimek, P., Vaněk, J., Jarolímek, J. Information and communication technologies and multifunctional agri-food systems in the Czech Republic. Plant, Soil and Environment, 2008, year. 54, Bo. 12, p. 547-551. ISSN: 1214-1178.

[4] Vaněk, J., Kánská, E., Jarolímek, J., Šimek, P. State and evaluation of information and communication technologies development in agricultural enterprises in Czech Republic. Plant, Soil and Environment, 2010, year. 56 (2010), No. 3, p. 143-147. ISSN: 1214-1178.

[5] Alexander K. Seewald, Wilfried N. Gansterer, On the detection and identification of botnets, Computers & Security, Volume 29, Issue 1, February 2010, Pages 45-58, ISSN 0167-4048, 10.1016/j.cose.2009.07.007.

[6] Xinyuan Wang, Daniel Ramsbrock, Chapter 8 - The Botnet Problem, In: John R. Vacca, Editor(s), Computer and Information Security Handbook, Morgan Kaufmann, Boston, 2009, Pages 119-132, ISBN 9780123743541, 10.1016/B978-0-12-374354-1.00008-X.

[7] Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali, Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification, Computer Networks, Volume 53, Issue 6, 23 April 2009, Pages 835-848, ISSN 1389-1286, 10.1016/j.comnet.2008.11.012.

[8] Gu-Hsin Lai, Chia-Mei Chen, Chi-Sung Laih, Tsuhan Chen, A collaborative anti-spam system, Expert Systems with Applications, Volume 36, Issue 3, Part 2, April 2009, Pages 6645-6653, ISSN 0957-4174, 10.1016/j.eswa.2008.08.075.

[9] Spam levels drop drastically … then rise, Computer Fraud & Security, Volume 2011, Issue 1, January 2011, Pages 1-3, ISSN 1361-3723, 10.1016/S1361-3723(11)70001-X.

[10] Authon original research

[11] Zhenhai Duan, Kartik Gopalan, Xin Yuan, An empirical study of behavioral characteristics of spamers: Findings and implications, Computer Communications, Volume 34, Issue 14, 1 September 2011, Pages 1764-1776, ISSN 0140-3664, 10.1016/j.comcom.2011.03.015.

[12] Tyrychtr, J., Buchtela, D., Havlicek, Z. Návrh ROLAP databáze v zemědělském podniku: Transformace ekonometrického modelu do konceptuálního modelu dat. Systémová integrace, 2012, year 19, No. 2, p. 51 - 61. ISSN: 1210-9479.