Hacking Malicious Users on Social Networks Using COMPA Method

Surya Suresh, Prof. Ushus Maria Joseph

Department of CSE, MBCCET, Kerala, India

ABSTRACT: Online social networks, such as Facebook and Twitter, have become one of the main media to stay in touch with the rest of the world. Celebrities use them to communicate with their fan base, corporations take advantage of them to promote their brands and have a direct connection to their customers. But today unfortunately control over many accounts fall into the hands of cyber criminals. Compromising social network accounts has become a profitable course of action for cyber criminals. By hijacking control of a popular media or business account, attackers can distribute their malicious messages or disseminate fake information to a large user base. The impacts of these incidents range from a tarnished reputation to multi-billion dollar monetary losses on financial markets. COMPA is an efficient method to identify compromises of individual high-profile accounts. High-profile accounts, frequently have one characteristic that makes this detection reliable, they show consistent behaviour over time. This paper consists of the literature survey on COMPA method to study the methods and features used to develop COMPA. Make a study for providing more security measures for the COMPA method.

KEYWORDS: COMPA, Spam detection

I. INTRODUCTION

A social networking service (also social networking site, or SNS or social media) is an online platform which people use to build social networks or social relations with other people who share similar personal or career interests, activities, backgrounds or real-life connections. Facebook and Twitter are two popular social media that celebrities use them to communicate with their fan base, corporations take advantage of them to promote their brands and have a direct connection to their customers. Giving people the power to share and make the world more open and connected. Unfortunately, the control over several accounts fall into the hands of a cyber criminal, thus he can easily exploit this trustworthy account to further his own malicious agenda. Previous research showed that using compromised accounts to spread malicious content is advantageous to cyber criminals, because social network users are more likely to react to messages coming from accounts they trust.

Traditional attacks such as sending spam messages or link to malware and phishing websites are mainly carried out using compromised attacks. An abundance of research was proposed in the most recent years to recognize malicious movement on online social communities. The vast majority of these frameworks, focus on detecting fake accounts specifically created to spread malicious content, instead of looking for legitimate accounts that have been compromised. To recognize compromised social network accounts this paper proposes a method called COMPA. The social account users develop behavioural patterns and it will be fairly stable. To detect account compromises, the system builds a behavioural profile for each social network accounts, based on the past activities. Every time a new activity is performed, it is compared against this behavioural profile. If the behaviour notably deviates from the learned behavioural profile, COMPA flags it as a possible compromised. COMPA can reliably detect compromises that affect high profile accounts. Since the behavior of these accounts is very consistent, false positives are minimal. High-profile accounts frequently have one characteristic that makes this detection reliable.

II. LITERATURE SURVEY

The following are the papers surveyed for need of doing the project to find out the unknown attacks on the social networking site. These papers make study about various attacks in ONSs and proposes different techniques to find those attacks and to detect the compromised accounts.

[1] A STUDY ON WEB SPAM CLASSIFICATION AND ALGORITHMS

Intentional attempt to manipulate search engine rankings for specific keywords or keyword phrase queries is considered as web spam. Facebook and Twitter are not immune to messages containing spam links. Most insidiously, spammers hack into accounts and send false links under the guise of a user's trusted contacts such as friends and family. As for Twitter, spammers gain credibility by following verified accounts such as that of Lady Gaga; when that account owner follows the spammer back, it legitimizes the spammer. Twitter has studied what interest structures allow their users to receive interesting tweets and avoid spam, despite the site using the broadcast model, in which all tweets from a user are broadcast to all followers of the user. Spammers, out of malicious intent, post either unwanted (or irrelevant) information or spread misinformation on social media platforms. Spreading beyond the centrally managed social networking platforms, user-generated content increasingly appears on business, government, and nonprofit websites worldwide. Fake accounts and comments planted by computers programmed to issue social spam can infiltrate these website. Web spam is classified into different types based on the impact of spam in various areas in internet. It is mainly classified as content spam, social network spam, email spam, image spam, click spam, cloaking and redirection and link spam.

1. Content spam

By definition, spam is unwanted, intrusive, and irrelevant advertising on the Internet. Spam content is content that is not created with the intention of serving your audience. When you steal other content in your website links is also considered as content spam. Based on the structure of a document content spam is classified into title spam, body spam, meta-tag spamming, anchor text spamming and URL spamming. Title spamming is the spamming the title of a document is a serious issue. The use of repeated words or phrases in the title is not permitted. In Body spamming the document body may contain spam words. It is the most common form of spamming and is old as the search engines themselves. Document header contains HTML meta-tag, which is always been the target of spamming. Spam terms are sometimes included in the anchor text of the HTML hyperlinks to a page. The URL will be made longer by adding spam words.

2. Social network spam

Social spam is unwanted spam content appearing on social networking services, social bookmarking sites, and any website with user-generated content (comments, chat, etc.). It can be manifested in many ways, including bulk messages, profanity, insults, hate speech, malicious links, fraudulent reviews, fake friends, and personally identifiable information.

3. Email spam

Email spam, also known as junk email, is unsolicited messages sent in bulk by email. Most email spam messages are commercial in nature. Whether commercial or not, many contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware.

4. Image spam

Image-based spam, or Image spam, for short, is a kind of E-mail spam where the textual spam message is embedded into images that are then attached to spam emails. Since most of the email clients will display the image file directly to the user, the spam message is conveyed as soon as the email is opened. The goal of image spam is clearly to circumvent the analysis of the email's textual content performed by most of the spam filters.

5. Click spam

Also known as organics poaching, click spam is a type of fraud which happens when a fraudster executes clicks for users who haven't made them. Click spamming captures organic traffic, brands it without its knowledge and then claims the credit for the user later.

6. Link spam

Link spam is the posting of out-of-context links on websites, discussion forums, blog comments, guest books or any other online venue that displays user comments. Link spam is also known as comment spam, blog spam or wiki spam. Link spammers usually don't leave comments of any value along with their links.

Content-based Spam Detection and Link –based Spam Detection are two algorithms that can be used for detecting the web spam. Content-based Spam Detection uses statistical analysis since spam pages are usually automatically generated, using phrase stitching and weaving techniques. There are five kinds of Link –based Spam Detection which uses exploits various techniques for spam detection.

[2] A SURVEY ON VULNERABLE ATTACKS IN ONLINE SOCIAL NETWORKS

Vulnerability is a weakness which can be exploited by a Threat Actor, such as an attacker, to perform unauthorized actions within a computer system. To exploit vulnerability an attacker must have at least one applicable tool or technique that can connect to a system weakness. In this frame, vulnerability is also known as the attack surface.

The online social networking faces an increased rate of security threats. The networking providers will always there to provide security but the attackers will tries to break their security measures by introducing vulnerabilities. Some attackers mine user's personal information for fun, the motive behind most of the attacks is harassment, steal personal information, company information, information related to bank accounts, security numbers and password will cause serious issues. The vulnerable attacks in online social networks are basically classified into three such as classic threats, modern threats and adolescent attacks.

1.Classic threats

From the introduction of the Internet, classic threats are there around. It mainly includes the installation of viruses into the target system. The following are some of the examples for classic threats:

- Malware: Malware, or malicious software, is any program or file that is harmful to a computer user. Malware includes computer viruses, worms, Trojan horses and spyware.
- Spammers: A spammer is a person or group that sends you email you don't want and didn't sign up for. The email a spammer sends is called spam. The words spammer and spam come from the canned meat called Spam and the Monty Python comedy sketch that was inspired by it.
- Phishing attacks: Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message.
- Cross-site scripting or XSS attack: Cross-Site Scripting (XSS) attacks are a type of injection, in which malicious scripts are injected into otherwise benign and trusted websites. XSS attacks occur when an attacker uses a web application to send malicious code, generally in the form of a browser side script, to a different end user.

2. Modern threats

The recently evolved threats that are mainly affects social networking is considered as modern threats. There are different kinds of threats under this category.

- Click jacking: Click jacking is a malicious technique of tricking a Web user into clicking on something different from what the user perceives they are clicking on, thus potentially revealing confidential information or taking control of their computer while clicking on seemingly innocuous web pages. It is a browser security issue that is a vulnerability across a variety of browsers and platforms. A click jack takes the form of embedded code or a script that can execute without the user's knowledge, such as clicking on a button that appears to perform another function. Click jacking is an instance of the confused deputy problem, a term used to describe when a computer is innocently fooled into misusing its authority.
- Socialbot: A socialbot is a type of bot that controls a social media account. Like all bots, a socialbot is automated software. The exact way a socialbot replicates depends on the social network, but unlike a regular bot, a socialbot spreads by convincing other users that the socialbot is a real person.
- Fake profiles: Fake profiles also known as Sybil attack. The Sybil attack in computer security is an attack where in a reputation system is subverted by forging identities in peer-to-peer networks. It is named after the subject of the book Sybil, a case study of a woman diagnosed with dissociative identity disorder.

3. Adolescent attacks

As the notoriety of the social communities is taking off high on cloud, the contribution of the youths in social networking websites is expanding each day. cyberbullying and online grooming are examples of this kind of attack.

- Cyberbullying: It is a form of bullying or harassment using electronic means. It is also known as online bullying. Cyberbullying is when someone, typically teens, bully or harass others on social media sites. Harmful bullying behavior can include posting rumors, threats, sexual remarks, a victims' personal information, or pejorative labels
- Online Grooming: Grooming is when someone builds an emotional connection with a child to gain their trust for the purposes of sexual abuse, sexual exploitation or trafficking.

[3] N-GRAM-BASED TEXT CATEGORIZATION

Text categorization is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world. It is the fundamental task in the document processing. Presence of textual errors makes the text categorization a difficult task. N-gram-based text categorization is an efficient text categorization method that is tolerant of textual errors. The system is small, fast and robust. In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. An N-gram is an N-character slice of a longer string. This system uses n grams of several various lengths at the same time. At the beginning and ending of each gram blanks are also added. The word "GRAM" would be composed of the following N-grams:

bi-grams: _ G,GR,RA,AM,M _
tri-grams: _ GR,GRA,RAM,AM _,M_ _
quad-grams: _GRA,GRAM,RAM _,AM _ _,M_ _ _ _

In general, a string of length k, padded with blanks, will have k+1 bi-grams, k+1tri-grams, k+1 quad-grams, and so on. The N gram based text categorization is based on Zipf's Law. The law states that,

"the nth most common word in a human language text occurs with a frequency inversely proportional to n".

The law is applied in the concept that according to frequency of use there is always some set of words that dominates most of the other words of the language.

GENERATING N-GRAM FREQUENCY PROFILES

Here the proposed system reads the incoming documents and counts the occurrences of all N grams. The following procedures will be done for each incoming text:

- Split the text into tokens consisting only of letters and apostrophes by discarding digits and punctuations.
- Scan down each tokens.
- Hash into a table to find the counter for the N-gram, and increment it.
- Determine all N grams and their respective count.
- Sort that count in decreasing order to get reverse N-gram frequency profile for the document.

The below figure shows the dataflow for the proposed system. The n gram based text categorization works highly efficient for text from noisy sources like email and OCR systems. This approach will perform word stemming automatically that is beneficial for text categorization. The propose approach has the ability to work equally well with short and long documents, and the minimal storage and computational requirements. It is highly inexpensive and effective way of text categorization.

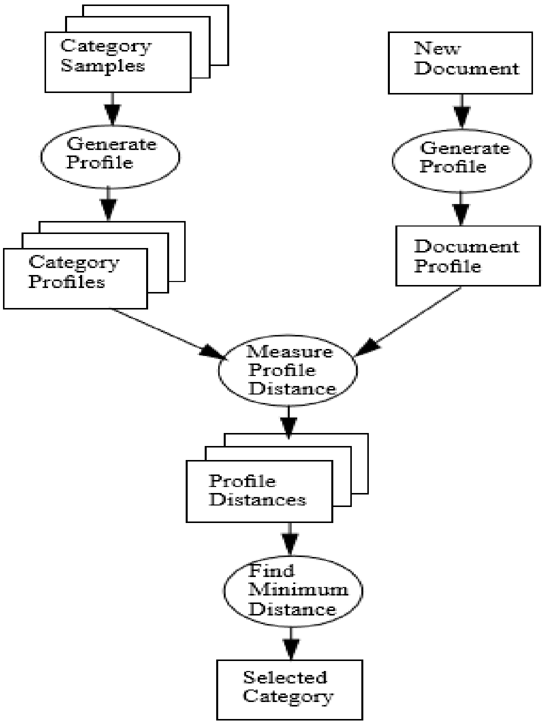

Fig.1.Dataflow for N-Gram-Based Text Categorization

[4] SURVEY PAPER FOR WARNINGBIRD: DETECTING SUSPICIOUS URLS IN TWITTER STREAM

A social networking service (also social networking site, or SNS or social media) is an online platform which people use to build social networks or social relations with other people who share similar personal or career interests, activities, backgrounds or real-life connections. There are several social networking sites and these shows tremendous growth in recent years. Social networking sites allow users to share ideas, digital photos and videos, posts, and to inform others about online or real-world activities and events with people in their network. This paper mainly focuses on Twitter and investigates the tweets that contain malicious URLs. Twitter limits its tweet length by 140 characters. A single URL may carry characters that exceed this limit. Thus people may use URL shortening services to reduce the length of URL before posting the tweet. t.co is the default URL shortening service of Twitter. Warning bird is the proposed suspicious URL detection system that investigates correlations of URL redirect chains extracted from several tweets in Twitter.

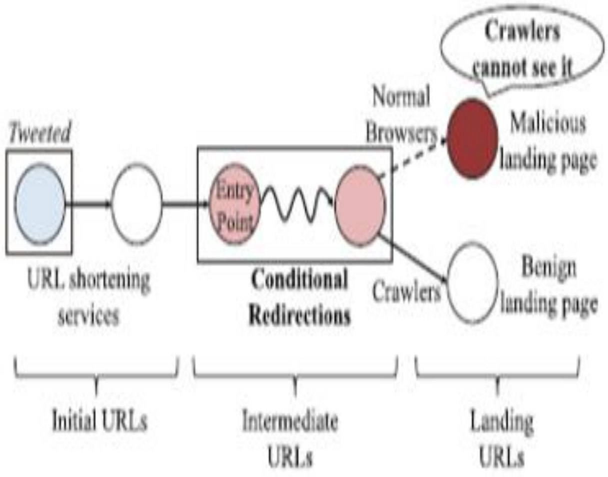

Fig.2.Conditional redirection

A malicious URL will contain several redirection URLs before landing to the malicious page. When a user or a crawler visits the initial URL, he will be redirected to an entry point of the intermediate URLs. The intermediate URLs are associated with private redirection servers which checks whether the visitor is a normal browser or a crawler by examining the IP address. The correlation analysis of warningbird can help to detect suspicious URLs even when they perform conditional redirection.

The suspicious URL detection of Warningbird is composed of four major phases such as:

- Data collection: It collects the tweets with URLs from Twitter dataset. This phase uses Twitter Streaming APIs to collect the data.
- Feature extraction: This includes subcomponents such as grouping identical domains, finding entry point URLs, and extracting feature vectors. This system maintains a tweet window that contains a particular number of tweets. Each tweet in the tweet window is analyzed and those share same IP address. This grouping process allows the detection of suspicious URLs that use several domain names to post malicious data.
- Training: Information about account statuses and a training classifier is required to train the collected data. The account statuses are used to label the training vectors.
- Classification: Executes classifier using input feature vectors to classify suspicious URLs.

It uses different features that related to correlated URL redirect chains and tweet context information for classification. URL redirect chain length, Frequency of entry point URL, Position of entry point URL, Number of different initial URLs, Number of different landing URLs are the features Derived from Correlated URL Redirect Chains. Number of different source, Number of different Twitter accounts, Standard deviation of account creation date, Standard deviation of the number of followers and number of friends, Standard deviation of the follower-friend ratio and Tweet text similarity are the Features Derived from Tweet Context Information.

[5] DETECTING SPAMMERS ON TWITTER

Twitter is one of the most popular social networking sites. There different kinds of real time search systems and mining tools for get the information about the consequences of events and news on Twitter. With the tremendous growth of Twitter the spammers are also increasing. When a significant event occur a large number of people will tweet about that event. These topics become the target of spammers that they include malicious URLs along with the tweet about the event. Since Twitter limits its tweet length users need to use URL shortening services while posting URLs.

The malicious URL may used to spread advertise to generate sales, disseminate pornography, viruses, phishing, or compromise system reputation. This makes difficult to identify the content in the URL without loading that page. Thus an efficient mechanism is required to reduce such attacks.

The proposed system contains three phases such as initial dataset and labeled collection, identifying user attributes and spam detection. Initially the system will make a labeled collection of users that is already classified as spammers and non spammers. It needs to crawl the Twitter to collect details about existing Twitter users including their social connections, and all the tweets they ever posted. Next it builds a labeled collection spammers and non spammers. For that the collected data set must contain spam users and non spam users. The spammers in the dataset must be aggressive in their strategies and it is desirable that users are chosen randomly and not based on their characteristic. Then the next step is identifying the user attributes. The spammers and legitimate users will always have different goals which make them behave different. Non-spammers spend more time interacting with other users, doing actions like replying, retweeting, posting status without URL, etc. Thus to identify the user characteristics the behavior of users in the labeled collection are examined. To analyze it a large set of user attributes are considered namely, content attributes and user behavior attributes. Content attributes are obtained from the text of tweet posted by each user. The metrics used for analyzing the content are the number of hash tags per number of words on each tweet, number of URLs per words, number of words of each tweet, number of characters of each tweet, number of URLs on each tweet, number of hash tags on each tweet, number of numeric characters that appear on the text, number of users mentioned on each tweet, number of times the tweet has been retweeted. The user behavior attributes are considered based on their interactions, number of posts, influence on the twitter network, number of followers, number of followees, fraction of followers per followees, number of tweets, age of the user account, number of times the user was mentioned, number of times the user was replied to, number of times the user replied someone, number of followees of the user's followers, number tweets received from followees and the minimum, maximum, average, and median of the time between tweets, number of tweets posted per day and per week.

The final phase is detecting spammers. It uses a supervised learning algorithm that learns a classification model from previously collected labeled data set. Thus the learned model can be applied to newly collected data to classify the users as spammers and non spammers. Support Vector Machine can be used as the classifier in the practical implementation. The experimental results show that, using this classification mechanism almost input data could be successfully classified, only ignorable portion is misclassified. It is find that using different subset of attributes, that the proposed classification detects spammers accurately.

[6]EVILCOHORT: DETECTING COMMUNITIES OF MALICIOUS ACCOUNTS ON ONLINE SERVICES

Social networking sites attained tremendous growth in recent years. Thus cyber attackers use the social accounts for malicious activities. They try to spread malicious content and to get confidential data from legitimate users. A botnet is a number of Internet-connected devices, each of which is running one or more bots. Botnets can be used to perform distributed denial-of-service attack (DDoS attack), steal data, send spam, and allows the attacker to access the device and its connection. Cyber criminals mainly exploit botnets for their malicious activity. The proposed system EVILCOHORT only requires the mapping between an online account and an IP address to detect any malicious activity that an account performs. There are several malicious campaigns in social media that do malicious activities. Such campaigns require social accounts and connection points to perform any malicious activity. All social media requires the creation of account and login to access the accounts. Connection points, which are the devices that run client software, are the mean through which attacker access social accounts. These connection points can also be bots. The proposed approach analyzes the connection between an attacker and an online service. The attackers will always use social accounts in a different manner than legitimate users. They may have many account and con nection point for an individual. The proposed system observes a number of IP addresses and accounts, and each account is accessed by a non-trivial portion of these IP addresses. These IP addresses will belong to bot infected system and attacker uses it to log onto the social accounts. The system keeps a record of interaction events of each account and the corresponding IP addresses that include the data about logging in, send mail and send friend request. Based on the data collected from interaction events the proposed system will build a bipartite graph where one set of one set of vertices is the online accounts observed and the other set of vertices is the list of IP addresses that accessed them. Then computes the weighted one mode projection of this graph to form projected graph representation. It uses a clustering method and iteration algorithm for filtering malicious communities from the projected graph representation. Legitimate communities behave differently from malicious communities and this deviation helps in the detection of illegitimate communities. The EVILCOHORT has been implemented on two real world datasets. The experimental results show that the proposed system detects malicious accounts very accurately.

[7] UNCOVERING SOCIAL SPAMMERS: SOCIAL HONEYPOTS + MACHINE LEARNING

In recent years community based interactions are provided by many web based social systems for the users. Social accounts prone to cyber attacks, that may cause serious issues for legitimate users. Traditional spam detection systems like email spam detection which doesn't harm the ongoing ability of the attacker. Inorder to uncover the spammers on the online social systems more efficiently here proposes a honey pot based approach. A honey pot is a computer system that is set up to act as a decoy to lure cyber attackers, and to detect, deflect or study attempts to gain unauthorized access to information systems. The primary features of the proposed Honey pot based approach are (1) Honey pot is used to collect the spam profiles from social systems (2) To filter new spammers in the online social systems a classifier is build by performing statistical analysis on the harvested data in the Honey pot. The Honey pots will monitor the behavior of spammers and log them. Thus it enables automatic detection of spammers. The presented system implements a Honey pot system that consist collection of profiles of spammers and non spammers. A bot is made associated with the Honey pot that detects social spam activities. It will find the evidences for malicious activity.

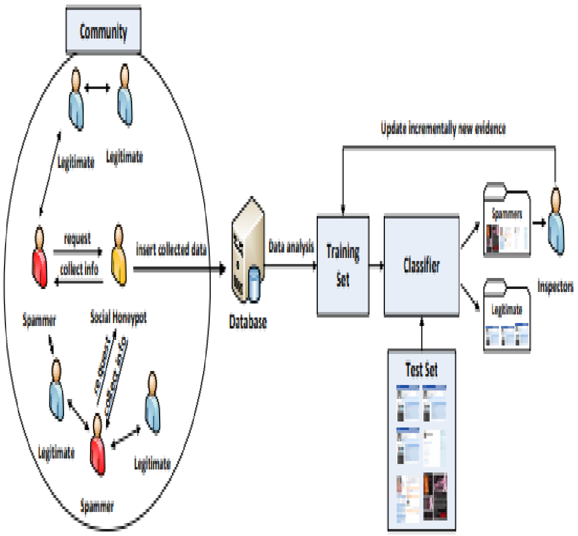

Fig.3. Overall Framework of Social Honeypot-based Approach

III. CONCLUSION

In the past there have been so many studies that are done to study the compromised account detection in social networks. Compromising social network accounts has become a profitable course of action for cyber criminals. By hijacking control of a popular media or business account, attackers can distribute their malicious messages or disseminate fake information to a large user base. The impacts of these incidents range from a tarnished reputation to multi-billion dollar monetary losses on financial markets. COMPA is an efficient method to identify compromises of individual high-profile accounts. High-profile accounts, frequently have one characteristic that makes this detection reliable, they show consistent behaviour over time. This paper consists of the literature survey on COMPA method to study the methods and features used to develop COMPA. Make a study for providing more security measures for the COMPA method.

REFERENCES

[1] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, "User interactions in social networks and their implications," in Proc. 26th Annu. Comput. Security Appl. Conf., 2010, pp. 11–20.

[2] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in Proc. 3rd Annu. Symp. Document Anal. Inform. Retrieval, 1994, pp. 161–175.

[3] S. Lee and J. Kim, "WarningBird: Detecting suspicious URLs in twitter stream," in Proc. Symp. Netw. Distrib. Syst. Security, 2012

[4] Z. Cai and C. Jermaine, "The latent community model for detecting sybils in social networks," in Proc. Symp. Netw. Distrib. Syst. Security, 2012, pp. 563–578

[5] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a twitter network," First Monday, vol. 15, no. 1, 2010.

[6] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2010, pp. 435–442.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in Proc. Conf. Email Anti-Spam, 2010, vol. 6, p. 12.