Surveillance Issues for Security over Computer Communications and Legal Implications

Mrs. Shilpa Mehta, Dr. U Eranna, Dr. K. Soundararajan

***Abstract*— This paper discusses computer communications security issues, which cannot be addressed by encryption and decryption alone. It points out the need for automatic electronic surveillance of mails without human intervention, and also the equally important need for educating the consumers about possibility of surveillance without infringing on their privacy rights. The need for redefining our legal systems and formulating new laws to combat this problem is also underlined. In today's situation of terabytes of traffic moving over the Internet, the detection of terrorism and crime related mails is a rapidly growing challenge. This is already being dealt with by backward learning and suspicious words' watch lists. Here these methods as well as methods of testing for word substitutions are explored.**

***Index Terms* — Computer Communications, Surveillance, Word Substitutions, Watch lists, Word Counts.**

I. Introduction

AS we are all aware, computer communications, especially e-mails, are very much in vogue today. This has definitely made life a lot more convenient for the common man, by allowing him to communicate with friends and family at very low (or even zero) cost, and that too with an almost zero delay. But, unfortunately, along with the good, comes the bad. The other side of this shining coin is that, subversive groups are using the very same communication facilities to plan and execute terrible acts.

While we deal with this subject, we repeatedly come across terms like Carnivore, FBI (Federal Bureau of Investigation), NSA (National Security Agency), CSS (Central Security Service), DoD (Department of Defense USA). The legal Acts and reforms like the FISA (Foreign Intelligence Surveillance Act), The PATRIOT Act (Provide Appropriate Tools Required to Intercept and Obstruct Terrorism), The Protect America Act, TIA (Total Information Awareness program and Terrorism Information Awareness), the IAO (Information Awareness Office) are also commonly used names. We discuss all these terms in detail in the next section. It is pointed out that, in view of the new threats posed due to the technological advancements (facilitated by the information revolution of the last century), the legal systems of most countries are ill-adapted to handle the new challenges. The law making agencies need to stir themselves up and form new laws, and that too at a brisk pace, to adapt and become capable of successfully combating these novel crimes. We shall deal with some of the legal issues associated with surveillance of communications in section II of this paper.

Technologically speaking, we discuss the measures and counter measures being applied for surveillance of communications. The technical issues of non-utility of encryption in these situations and the primary technique of keyword filtering with its limitations are discussed in section III of this paper. The countermeasures of text substitutions and technical measures being developed and used to detect such substitutions are inspected in the IV section. Extended measures for detecting substitutions are dealt with in the V section of this paper.

Manuscript received February 7, 2010.
Mrs. Shilpa Mehta is an Assistant Professor in the department of Instrumentation Technology at PDIT Hospet, VTU Belgaum India. Her Phone No. is +91 08394 221050. and her email address is shilpartha@yahoo.com.
Dr. U Eranna is the Professor and Head of the department of Electronics and Communication Engineering at BITM Bellary, VTU Belgaum, India. His email address is jayaveer_88@yahoo.com.
Dr. K. Soundararajan is the Rector of JNTU Anantapur, India. His email address is soundararajan_jntucea@yahoo.com

II. The Legal Situation

As long back as 2000, the Carnivore[1] was being used by the FBI in the USA. The Carnivore system was a packet-sniffing workstation, and had to be physically installed at an ISP or other location where it could "sniff" traffic on a LAN segment to look for email messages in transit. The government personnel were required to get a warrant or court order naming specific people or email addresses to be monitored. When an email matching the filtering criteria of the warrant passed through, the message was logged along with information on the date, time, origin and destination. This logging was believed to be relayed in real time to the FBI. All other traffic would presumably be dropped without logging or capture. Following huge controversies, the name was changed to DCS1000 [DCS standing for Digital Collection System]. Despite this, the functioning of the system remained largely unchanged.

After the September 11 attacks in the USA in 2001, serious legal and technical measures were initiated to prevent the repetition of such incidences in future. The NSA/CSS [2] which is the US government's cryptographic intelligence agency (administered by the US-DoD), came up with TIA (Total Information Awareness program), to do extensive *warrant less* data-mining to identify "information signatures" which could identify criminals. The TIA was as system level program of the IAO (Information Awareness Office established by the Defense Advanced Research Projects Agency (DARPA)) for integrating information technologies into a prototype system, to provide tools to better detect, classify, and identify potential foreign terrorists. The goal

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

was the creation of a "counterterrorism information architecture" that integrated technologies from other IAO programs (and elsewhere, as required). After a public outcry, the department renamed it Terrorism Information Awareness. Now, the NSA has shifted to e-mail, Web and data-mining dragnet [3].

There are also reports [4] of Surveillance of Skype Messages in China. The full text chat messages of TOM-Skype users, along with Skype users who have communicated with TOM-Skype users, are regularly scanned for sensitive keywords, and if such words are found to be present in the communication text, the resulting data are uploaded and stored on servers in China. These text messages, along with millions of records containing personal information, are stored on insecure publicly-accessible web servers together with the encryption key required to decrypt the data. The captured messages contain specific keywords relating to sensitive political topics such as Taiwan independence, the Falun Gong, and political opposition to the Communist Party of China.

The FISA-1978 [5] (Foreign Intelligence Surveillance Act) was initially targeted at monitoring suspected espionage communications by spy networks, which were mainly working for other governments. This act was amended in 2001 by the USA PATRIOT act, primarily to include terrorism and to monitor the activities done on behalf of *groups not backed by a foreign government.* This was followed by the Protect America Act in 2007 and FISA Amendment Act in 2008. But even the FISA is not sufficiently technically competent to counteract the latest threats. Additionally, all countries do not have sufficient laws to cover the rising threats posed by the Internet. As we are all aware, Internet crime is rising to alarming levels. Laws need to be amended and new laws need to be created to handle this new tidal wave of danger.

Despite the public opposition, electronic surveillance is rapidly spreading to the overlapping areas of crime, terrorism and contemporary warfare [6]. Many steps have been taken, and are being taken today to address the requirements to protect the people from the increasing dangers of cyber crime. The laws and the surveillance are not meant to annoy or trouble innocent citizens, but to save their life and protect their property. People need to be educated and re-educated in this matter. As most democratic countries' constitutions stress on the freedom of the individual, the freedom of speech, communication and privacy, there is a ludicrous opposition to the proposed surveillance of Internet Communications. The very same citizens, who don't even bat an eyelid when they are scanned at airports or shopping malls, take great exception to their mails being passed through filters. People need to be reminded that even though email has arrived in full strength today, *the age old mail system is still working*. If they really want to send some highly confidential documents, they can use the established hardware sending system. But in the interests of the greatest good to the greatest numbers, for our own safety and security, we should not blindly oppose any kind of surveillance being used for scanning the millions of mails moving around in apparently innocuous packets over the Internet today. Yet, in spite of the closer monitoring of individuals, groups, governments, and states, our safety is not guaranteed. We need both better technology and better laws to combat the growing threats.

Moreover, people need to know that their mails are not being *read* by other people, just scanned by *machines* for sensitive words etc. They need reassurance that they will not be trapped into getting convicted for mails which accidentally fell into the net, or even forged emails! The conviction in the court of law must come only when sufficient physical proof (apart from the proof from electronic communication) is presented before the court.

In the age of older postal systems, the addresses were known and open to everyone, so senders and receivers were visible. We, the citizens, need to cooperate with our governments to ensure the same level of transparency and clarity in case of computer communications also; not treat it as a means of anonymous communication which is already illegal in most nations. On the other hand, people also need the reassurance from their lawmakers that these scans will only be used to narrow down the search criteria, and not as a harassment tool to trouble innocents! In the present scenario, it would be advisable for the governments also to initially introduce the concept as only an investigation tool, not as a evidential element in courts of law. If further investigations don't test positive, the surveillance should be dropped, without a lot of red tape. If governments take care of these factors while making laws, there is no reason for the educated and responsible population not to cooperate (and the uneducated ones would not be using the computer communication facilities anyway!).

The issues involved are similar to Internet censorship. Internet censorship involves control or suppression of either the publishing or accessing of information on the Internet. The legal issues are similar to offline censorship. One difference is that national borders are more permeable online. If a country bans certain information, residents can find it on websites hosted outside the country. A government has no control over the websites themselves, but it can try to prevent its citizens from viewing these. Filtering can be based on a blacklist or be dynamic. In the case of a blacklist, that list is usually not published. The list may be produced manually or automatically.

III. Technical Issues And the Limitations of Earlier Surveillance Techniques

Normally people feel that encryption is the only method used for hiding any communications from surveillance. They thus tend to concentrate on just decryption techniques applied to the intercepted communications. But things are not so simple. In actual fact, there are millions of emails moving around the Internet everyday. In such large numbers, it is very easy for subversive groups to send their illicit communications without encryption and the chances of detection are very rare. In fact, in such cases, the senders are well aware of the fact that encryption *may actually attract attention* to a mail which might otherwise have passed unnoticed. Hence, such mails are hardly ever encrypted, and decryption techniques are completely useless in this scenario.

During earlier days of surveillance of communications, Keyword filtering to find significant words ('attack' or 'bomb') was used for detecting suspicious communications. It works like the Spam filters used by most email serving applications. But the subversives soon started using word substitutions to fool this system also. This kept the authorities fooled for a long time as they were only searching by lists of

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

suspicious words. Now, techniques based on language and word usages are being devised to detect such text word substitutions. But we have to keep in mind that fighting this threat is like fighting an ever mutating indestructible foe, and we should be constantly on our toes and on the lookout to keep this enemy at bay. We shall discuss measures to detect such substitutions in the next section.

IV. The Surveillance and Counter Surveillance Techniques in use today

In the earlier days of electronic communications, it was imagined that encryption was the answer to all kinds of unwanted probes into the intended private communication. But, just like RADARS in warfare, no measure is a perfect solution to all problem domains. Measures, Counter measures and Counter-counter measures are being invented every day. People who want to hide their message contents *know* that systems like Echelon are expected to intercept their communications.

Echelon [7] is a satellite-based system designed to monitor almost any kind of electronic signals. It intercepts the signal, then a computer network implements software using keyword recognition and voice recognition to pick out sensitive words. Phone calls made on land lines, faxes, and emails are all susceptible to this powerful surveillance tool. A coalition of nations known as the UKUSA, comprised of the United States, the United Kingdom, Canada, Australia, and New Zealand, use the system, and these countries then exchange information about the intercepted communications. This year, Intelligence specialists have suggested that members of UKUSA seize millions of transmissions every hour.

One way to conceal content is to encrypt the messages, but there are many drawbacks. The first drawback is that encryption attracts attention in a normal pool of messages which are *not* encrypted. Secondly, it is also difficult to use in 'on the fly' communication methods like mobile phones. Additionally, nobody is very sure about the robustness of encryption as security agencies have stout decryption capabilities, whose strength is unknown to common people.

The agencies widely use Keyword Filtering [8]. Words like 'Attack", "Bombing", "Nuclear" etc are on the watch lists of mail scanning soft wares. Keyword filtering (a common algorithm in most spam filtering soft wares) is a well known technique used to block or intercept communications using such words. It is used to select messages requiring further scrutiny from a set of intercepted messages. The countermeasure used commonly by criminal groups is to replace this type of words by other innocuous words. But, as discussed below, this gives rise to unnatural sounding sentences. These unusual messages can be readily detected and also sets of such related messages can be detected as conversations, even when the end points have been purposely obscured by using stolen mobiles or temporary mail ids.

We are all familiar with "spell-check" in Microsoft word and other text processing soft wares. The algorithms like BaySpell, WinSpell, [9] and numerous others work in basically two ways. Firstly, the algorithm checks for spelling mistakes creating 'nonsense' words (e.g. *'frist'* in place of *'first'* is indicated as a wrong word and a suggestion is made to change it to *'first'*). Sometimes it so happens that the errors lead to valid words (e.g. causal for casual). These mistakes can also be easily identified as the wrong word does not fit in the context of the phrase or sentence. (this is a *causal* signal versus this is a *casual* meeting) Sometimes the phrase also may sound correct, but the sentence doesn't make sense. For instance, both the phrases 'piece of mind' and 'peace of mind' are correct, but the sentences 'The wife gave her husband and his girlfriend a *piece of mind*' and 'When I go to church I find *peace of mind*' could not be found properly meaningful if the phrases were to be interchanged. Hence, the sentence oddity can be used as a measure to detect such word substitutions. Similar techniques are employed to combat the checking of communication with substituted words in text.

To combat this kind of detections, the subversive groups have made lists of substitute words which result in meaningful sentences [10]. For instance, at one time, the word 'wedding' was being used for 'attack'. The sentence 'The *attack* is on this date and time at this location' was replaced by an equally meaningful sentence 'The *wedding* is on this date and time at this location'. As both events are similar in terms of the planning and organization requirements as well as sentence usage, such sentences sound perfectly meaningful to humans. There is no sentence oddity factor to help agencies to detect this type of word substitutions in text. Fortunately, methods do exist to detect such substitutions also. This is because the replaced words and the original words have their own associated usage frequencies in the language [11]. In the above example, 'Attack' is the 1072th common word in English while 'wedding' is the 2912th one. This discrepancy in usage will lead to easy detection of the altered message.

However, the replacement can be done by words of similar frequencies. For instance the words *bomb* and *alcohol* have frequencies of 3155 and 3154, which are so close as to not permit the detection based on frequencies. Here, the sentence oddity comes into play. While the sentence 'The bomb is in position' is a normal sentence, the sentence 'The alcohol is in position' definitely sounds awkward and gets trapped. Moreover, it is almost impossible to do such perfectly matching substitutions 'on the fly' and under stress, without the help of manuals for guidance.

Thus we can understand that there are many methods in use. While individual methods are not foolproof, a set of measures working together make it possible to detect unwanted mail communications. At the same time, innocent citizens do not need to worry, as the probabilities of normal conversations falling into these traps are astronomical!!!

V. The Extended Measures

As we find in the commonly used search engines such as Google search, the words are capable of being subjected to individual searches [8],[9]. They each have their own individual identity apart from the role that they play in sentences. At the same time, we have to remember that the sentences being used are a particular set of words arranged in a particular fashion. But apart from being arranged in a well defined way, the words are also individual players and sentences may be treated as 'bags' of words. If the 'suspect' word is removed or replaced, the remaining 'bag' would have a much lower frequency of occurrence than the original one.

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

This supports the earlier intuitive conclusion that we observed in the sentences discussed in the example cases considered in the previous section.

The sentence oddity measures works as follows:

Let

f_{wo} = Frequency of occurrence of bag of words with suspect word removed

f = Normal frequency of occurrence of the bag of words

Then sentence Oddity is defined as

$$SO = f_{wo} / f \quad \ldots (1)$$

.Hence, sentence oddity will be larger for sentences containing substitutions than the original sentences. Similar to this is the Enhanced Sentence Oddity measure is defined as

$$ESO = f_{wo}' / f \quad \ldots(2)$$

Where f_{wo}' is the frequency with the target word specifically excluded (while f_{wo} could have accepted some repetitions' of the denominator also).

Another common type of substitution used is to replace some noun by its hypernym. One chain is "kitty (frequency 13131); house cat ; cat (frequency 2532); feline (frequency 23711); carnivore (frequency 41906); eutherian mammal." Another chain with the corresponding word frequencies is as shown in Table 1.

TABLE I

HYPERNYM CHAIN FOR THE NOUN DOG

S.No	Class	Frequency
1.	ENTITY	6355
2.	ORGANISM	8340
3.	ANIMAL	1521
4.	CHORDATE	21502
5.	VERTEBRATE	27840
6.	MAMMAL	16075
7.	PLACENTAL	34335
8.	CARNIVORE	41906
9.	CANINE	20227
10.	DOG	1279

Once again such substitutions will also create an oddity measure to correspond to the awkwardness introduced due to this substituted hypernym. For instance, if we look at the sentences '*John took Timmy for a walk*' or '*The man took his dog for a walk*' and compare these to '*The human being took his carnivore /mammal for a walk.*', it illustrates to us how odd the substituted sentences sound to human readers. This oddity measure is defined as

$$Oddity = HO = f_H - f; \quad \ldots(3)$$

Here, f_H and f are the frequencies of occurrence of the sentence with the generalized form and the original noun respectively. Other measures like left K-gram, right K-gram and PMI are also used with a satisfactory rate of success.

K-grams [12] are measures of frequencies for strings of limited length. The left k-gram of a word is defined as the string beginning at the word and extending left till the first non-stopword. Similarly right k-gram starts at the word and continues right till the first non-stopword.

We have already seen that, the frequencies of the fragments of sentences containing substitutions are lower than the original sentence. But, we do not know the frequency of the original sentence, because most quoted strings are of variable lengths and might even have never been seen before. Hence the reference values for comparison are not known. The frequencies of the exact strings are too low to work with (or even zero at times, when the original quoted string has never been seen before), and the comparison is not possible. The measure of k-gram frequencies is useful for building patterns whose natural frequencies can be practically estimated.

PMI [13] is the acronym used for Point wise Mutual Information. It is a measure for association and is used in Information theory and Statistics. The PMI of a pair of variables is a measure of the discrepancy between the probabilities of their coincidence (given their joint probability distribution) against the probability of their coincidence (given only individual distributions assuming independence). The defining Equation is

$$PMI(x,y) = \log [\{p(x,y)\} / \{p(x).p(y)\}] \quad \ldots(4)$$

Here, p(event) denotes the probability of the occurrence of that event. While studying intercepted communications, if we consider measuring the strength of association between a word of interest and an adjacent region of the sentence,

$$PMI = \{p(word) . p(adjacent\ region)\} \ / \ p(word + Adjacent\ region) \quad \ldots(5)$$

Here, the adjacent region may be in either direction. (i.e. the phrase following or preceding the word). The probabilities can be approximated by the inverse frequencies of occurrence of the word and the phrase respectively.

VI. CONCLUSION:

Currently, a debate is raging whether or not the president of the USA has the power to authorize the NSA to monitor communications. Nevertheless, even the people who are against this concept of surveillance, do agree that there is a need to intercept communications which pose a serious threat, and also to identify and monitor such conversations to enable the security agencies to take steps before the planned terrible attacks are successfully carried out [14] - [17] . Data mining technology provide us tools that can help us to trap such communications. Despite all this, we have to accept that the present status of legal structures as well as technologies, is inadequate and ill adapted to tackle all the new challenges posed before us by the new wave of criminal planning fuelled by the self same technology advancement. This paper is a step towards combating this threat which is unparalleled in history because it is a novel threat made possible by the technological developments of the information era.

REFERENCES

[1] Juri Stratford, "Internet Surveillance: Recent U.S. Developments" *IASSIST Quarterly, Fall 2003.*

[2] http://en.wikipedia.org/wiki/ National Security Agency, Central Security Service.

[3] Declan McCullagh "NSA shifts to e-mail, Web, data-mining dragnet", C NET news, Politics and Law, March 11, 2008.

[4] John Markoff, "Surveillance of Skype Messages Found in China" The New York Times, Oct 1 2008.

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[5] "Wikipedia", Foreign Intelligence Surveillance Act of 1978.
[6] Kirstie Ball and Frank Webster, "The Intensification of Surveillance: Crime, Terrorism and Warfare in the Information Age" Pluto Press, London, 2003.
[7] Bob Whitehead, "ECHELON surveillance system -- security or invasion of privacy?" http://www.video-surveillance-guide.com, August 26, 2005
[8] Fong, S., Skillicorn, D.B, and Roussinov, D., "Measures to Detect Word Substitution in Intercepted Communication" Proceedings of IEEE Intelligence and Security Informatics Conference (ISI 2006), May 23-24, 2006, San Diego, California.
[9] A R Golding and Dan Roth, "A Winnow-Based Approach to Context-Sensitive Spelling Correction" Machine Learning Journal, Springer Netherlands, Volume 34, number 1-February 1999;
[10] SzeWang Fong; Roussinov, D.; Skillicorn, D.B., "Detecting Word Substitutions in Text" IEEE Transactions on Knowledge and Data Engineering, Volume 20, Issue 8, Aug 2008.
[11] www.wordcount.org/main.php
[12] David Skillicorn, "Knowledge discovery for Counterterrorism and Law Enforcement", CRC Press, page 256.
[13] http://en.wikipedia.org/wiki/Pointwise_mutual_information
[14] K.A. Taipale, "Whispering Wires and Warrantless Wiretaps : Data Mining and Foreign Intelligence Surveillance" N.Y.U. rev. l. & security, no. vii supl. bull. on l. & sec. (Spring 2006) http://whisperingwires.info/
[15] John R. Schmidt (associate attorney general in the Justice Department under President Bill Clinton), A historical solution to the Bush spying issue, CHIC. TRIB. (Feb. 12, 2006)
[16] Arlen Specter, "The Need to Roll Back Presidential Power Grabs" The New York Review of Books, Volume 56, Number 8 · May 14, 2009
[17] "Transcript: Senator Dick Durbin on Surveillance, Terrorism and the Constitution" Newstin Oct 5 2009.

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)