International Journal of Performability Engineering, Vol. 8, No. 2, March 2012, pp.173-182.

Printed in India

Long PN Code Based Traceback in Wireless Networks

XIAN PAN[1], JUNWEI HUANG[1], ZHEN LING[2], BIN LU[3], and XINWEN FU[1]

[1]*University of Massachusetts Lowell, Lowell, MA 01854, U.S.A.*
[2]*Southeast University, China*
[3]*West Chester University, West Chester, PA 19383, U.S.A.*

(Received on October 1, 2010, revised on March 25, 2011)

Abstract: Cyber criminals may abuse open wireless networks or those with weak encryption for cyber crimes. To locate such criminals, law enforcement has to first identify which mobile (MAC) is generating suspect traffic behind a wireless router. The challenge is how to correlate the private wireless traffic and the identified suspect public traffic on the Internet. In this paper, we propose a new technique called long Pseudo-Noise (PN) code based Direct Sequence Spread Spectrum (DSSS) flow marking technique for invisibly tracing suspect anonymous wireless flows. In this technique, a long PN code is shared by two investigators, *interferer* and *sniffer*. Different bits of the signal will be encoded with different segments of the long PN code. By interfering with a sender's traffic and marginally varying its rate, interferer can embed a secret spread spectrum signal into the sender's traffic. By tracing where the embedded signal goes, sniffer can trace the sender and receiver of the suspect flow despite the use of anonymous encrypted wireless networks. Traffic embedded with long PN code modulated watermarks is much harder to detect. We have conducted extensive analysis and experiments to show the effectiveness of this new technique. We are able to prove that existing detection approaches cannot detect the long PN code modulated traffic. The technique is generic and has broad usage.

Keywords*: Anonymous traceback, DSSS, long PN code, wireless network*

1 Introduction

The number of cyber crimes has also been increasing drastically with the converged wireless networks and Internet. Cyber criminals can utilize open wireless networks, or easily hack the weak protected WiFi routers, get the Internet access and commit crimes. These crimes include sexual exploitation of children, intellectual property theft, identity theft, financial fraud, espionage, and many others.

The challenge of conducting cyber crime scene investigations in wireless networks is how to correlate the private wireless traffic and the identified suspect public traffic on the Internet because of the use of NAT (network address translation) in wireless routers. The suspect public traffic can be network attacking traffic or child pornography downloading traffic that has been identified by intrusion detection systems and Internet surveillance tools. Traffic correlation in unencrypted wireless networks is straightforward by packet ID and other traffic features. Traceback in encrypted wireless networks is complicated since encryption erases recognizable IP packet content. Once the private wireless traffic and the mobile MAC have been identified, further approaches such as 3DLoc [13] can be applied to locate the suspect for search warrant from courts

In this paper, we developed a new flow marking technique called *long PN code based DSSS watermarking* for invisible traceback and apply this new technique to wireless networks. In this technique, a long PN code is shared by two investigators (interferer and sniffer). The long PN code is used to spread a signal. One segment of the long PN code is

*Corresponding author's email: xinwenfu@gmail.com

used to spread one bit of the signal. Different bits of the signal will be encoded with different segments of the long PN code. Basically, we are using different codes to spread different signal bits. This defeats MSAC (the mean square autocorrelation attack) based detection in [3], which can only detect a spread signal with the same short PN code spreading all the signal bits.

We have conducted extensive analysis and experiments to show the effectiveness of this new technique. We are able to prove that MSAC based detection cannot detect the long PN code modulated traffic. We developed a suite of tools and performed real-world Internet experiments over encrypted wireless networks plus Anonymizer [1], which is a popular commercial anonymous communication network. Our data validate the theory and demonstrate that our long PN code based DSSS watermarking technique can invisibly trace anonymous traffic flow over encrypted wireless networks.

The rest of the paper is organized as follows. In Section 2, we briefly review the most related work. In Section 3, we introduce the long PN code based traceback. We analyze the benefits of the long PN code based traceback in Section 3.4. The real-world experimental results are presented in Section 4. We conclude this paper in Section 5.

2 Related Work

There has been much research on degrading anonymous communication through mix networks. Because of the space limit, we give brief review of most related work.

To determine whether Alice is communicating with Bob, through a mix network, similarity between Alice's outbound traffic and Bob's inbound traffic may be measured. For example, Zhu *et al*. in [4] proposed the scheme of using mutual information for the similarity measurement. Levine *et al*. in [5] utilized a cross correlation technique. Murdoch *et al*. in [6] also investigated the timing based threats on Tor [17] by using some compromised Tor nodes. Fu *et al*. [7] studied a flow marking scheme. Overlier *et al*. [8] studied a scheme using one compromised mix node to identify the "hidden server" anonymized by Tor.

Yu *et al*. [2] proposed a direct sequence spread spectrum (DSSS) based traceback technique, which could be maliciously used to trace users of an anonymous communication network. However, this short PN code based traceback approach is subject to the mean square autocorrelation attack in [3]. This paper addresses this issue via the long PN code, which is also able to well support parallel traceback because of abundant number of long PN codes and their long length.

Zhang *et al*. [15] proposed using multiple orthogonal PN codes to spread different watermark bits and embed them in randomly selected intervals. In practice, the number of orthogonal PN codes is limited. A long PN code proposed in this paper addresses this issue of scarceness of orthogonal PN codes. Random intervals can also be inserted into long PN code modulated traffic to further improve its effectiveness against detection, including the multi-flow attack in [9]. Zhang, Luo and Yang [16] used PN codes to modulate the packet inter-arrival times in order to embed a secret signal into the target traffic. We modulate traffic rate instead of packet inter-arrival times. Both strategies have pros and cons. We leave the comparison as our future work.

3 Long PN Code Based DSSS Based Traceback

In this section, we will first define the problem, and introduce our basic idea. We then discuss the long PN code. At last, we introduce the flow marking process of embedding a long PN code spread signal into suspect traffic and recovering it.

3.1 Problem Definition and Basic Idea

Figure 1 illustrates the forensic case we are studying. A suspect sender is communicating anonymously with a suspect receiver through an encrypted wireless network and Anonymizer [1], which is a popular commercial anonymous communication network. The use of Anonymizer will make the traceback via wireless networks more challenging. For example, the suspect receiver could be a criminal downloading prohibited content from an illegal server, i.e., suspect sender. The suspect traffic is identified. The problem is: how can the law enforcement manipulate the suspect traffic in order to confirm it is the suspect sender who is communicating with the suspect receiver.

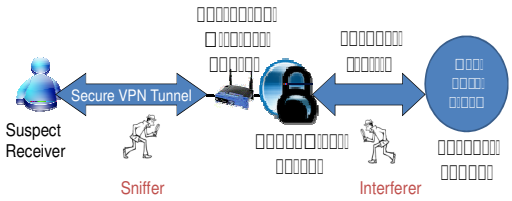


Figure 1: Encrypted Wireless Networks plus Anonymizer

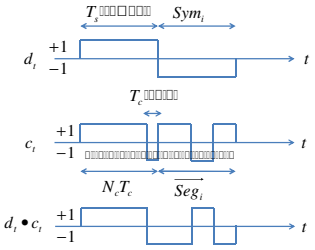


Figure 2: Long PN Code

Our basic idea to solve the problem is that if law enforcement *interferer* embeds a signal into the suspect traffic and law enforcement *sniffer* can recover the signal from the inbound traffic into suspect receiver, law enforcement confirms suspect sender communicates with suspect receiver. Techniques developed for this problem can be easily extended to a more general case: law enforcement can follow the traffic embedded with the signal and reconstruct the full communication path.

3.2 Long PN Code

In *Direct Sequence Spread Spectrum* (DSSS), we use *Pseudo-Noise* (*PN*) code to spread a signal over a bandwidth greater than the original signal bandwidth. Based on the length, there are short PN code and long PN code. In spreading and despreading processes, the two types of PN codes are very different. In short PN code based DSSS, the same short PN code is used to spread (encode) each bit of a signal.

Figure 2 shows the long PN code based DSSS technique, in which we use different segments of the long PN code to spread different signal bits. The original signal d_t is a series of binary symbols *Sym* (+1 or -1). The symbol duration for both symbol +1 and -1 is

(*MSRG*). We use MSRG to generate a long PN code. The configuration of a MSRG is determined by the primitive polynomial coefficients [10]. In Figure 3 the primitive polynomial is

$$f(x) = 1 + c_1 x + c_2 x^2 + \ldots + c_i x^i + \ldots + c_{n-1} x^{n-1} + x^n \quad (1)$$

where c_i is the coefficient, $i \in [1, n]$. c_i is either 0 or 1. R_i is the stage of the shift register. $\oplus$ refers to *XOR*. Different primitive polynomials generate different long PN

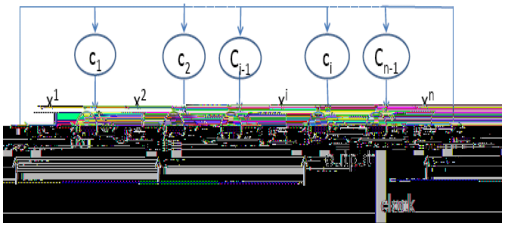


Figure 3: MSRG

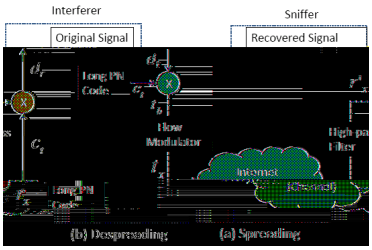


Figure 4: Long PN Code Spreading and Despreading in DSSS

codes. If the degree of the primitive polynomial is n, the number of different primitive polynomials of degree n is equal to the number of different long PN codes. The total number of different PN codes produced by primitive polynomials of degree n can be calculated as follows [10],

$$\text{Number of different long PN codes} = \phi(2^n - 1)/n \quad (2)$$

where $\phi(2^n - 1)$ is the *Euler's* ϕ *function*.

3.3 Flow Marking

Figure 4 illustrates the framework of flow marking. We spread a signal d_t as follows,

$$t_b = d_t \cdot c_t \quad (3)$$

where c_t is a segment of a partial long PN code and $\cdot$ is the element-wise multiplication of two vectors. t_b is then used to modulate a target traffic flow by interferer. We use *weak interference* against the flow when a chip is +1, so that the flow has a high rate for T_c seconds. We use *strong interference* against the flow when a chip is -1, so that the flow has a low rate for T_c seconds. We assume that the flow has an average traffic rate of D, then the high rate is $D+A$ and the low rate is $D-A$, where A is denoted as *mark amplitude*. The target traffic flow rate should be large enough for investigators to introduce the marks by interference. Therefore, the transmitted signal t_x can be represented by,

$$t_x = A d_t . c_t + D \quad (4)$$

The modulated flow travels through the Internet (including local WLAN and Anonymizer), where there exists noise created by cross traffic and other interference. We treat all noise n as an aggregated factor. So the received signal r_x is

$$r_x = A d_t . c_t + D + n \quad (5)$$

At the sniffer side (suspect receiver in Figure 1), in order to remove the direct current

component D from the received signal, a high-pass filter is applied. Therefore, the filtered received signal r'_x can be represented by,

$$r'_x \approx Ad_t.c_t + n \tag{6}$$

We then use the same segment c_r of the shared partial long PN code to despread the filtered received signal r'_x to derive the received baseband signal d_r,

$$d_r = Ad_t.c_t \cdot c_r + n \cdot c_r \tag{7}$$

A low-pass filter is then used to filter the high frequency noise. Thus,

$$d_r \approx Ad_t.c_t \cdot c_r \tag{8}$$

Since both *interferer* and *sniffer* have the same partial long PN code and $c_r = c_t$, $c_r \cdot c_t = 1$, we can recover the original signal.

3.4 Benefits of Long PN Code Based DSSS Based Traceback

In this paper, long PN code is applied in DSSS-based technique for tracing traffic flows in an anonymous network. By using long PN code, we can defeat *mean-square autocorrelation* (*MSAC*) based detection technique proposed in [3] and make the traceback hard to detect.

In this section, we will first present the partial correlation of the long PN code, then analyze the invisibility of the long PN code based-DSSS watermarking.

3.4.1 Partial Correlation of Long PN Code

Assume a long PN code is $C = \{c_0, c_1, \ldots, c_{P-1}\}$, where $c_i \in \{+1, -1\}$. The code period is P. A partial long PN code of length M from the whole long PN code is given by $C_s = \{c_s, c_{s+1}, \ldots, c_{s+M-1}\}$, where $s \in \{0, P-M\}$ and s is the starting position to get a segment of M chips from the long PN code. We calculate the correlation on the partial PN code C_s as follows,

$$r_{C_s}(\gamma) = \sum_{i=0}^{M-\gamma-1} (c_{i+s} * c_{i+s+\gamma}) \tag{9}$$

where $M < P$ and γ is the lag.

The mean value of the partial correlation for the PN code is presented in Lemma 1. The detailed proof of Lemma 1 is available in Appendix A of our technical report [14].

Lemma 1: $E\{r_{C_a}(\gamma)\}$ shows the mean value of the partial correlation, and γ is lag.

$$E\{r_{C_a}(\gamma)\} = \begin{cases} M, \gamma = 0 \\ -\dfrac{M-\gamma}{P}, \gamma \neq 0 \end{cases} \tag{10}$$

3.4.2 Invisibility of Long PN Code Based DSSS Based Traceback

The long PN code based DSSS watermarking technique makes it difficult to detect the fact of traceback by a suspect (receiver) being traced. A long PN code modulated traffic flow shows white noise-like pattern in both frequency and time domain. Suspects cannot detect those watermarks in frequency and time domains. The *mean-square autocorrelation* (*MSAC*) method also fails to detect the watermarks. The MSAC method is based on the fact that the same short PN code is repeatedly used to spread each signal bit. In our new technique, each bit is spread by successive different segments from a long PN

code. Basically, different signal bits are spread by different codes.

We now prove the invisibility of the long PN code based DSSS watermarking technique that can defeat the MSAC detection method. Denote $\overrightarrow{X} = x_0, \ldots, x_{N-1}$ as the signal, where N is the number of signal bits. x_i is either A or $-A$, where A is the watermark amplitude. Denote $\overrightarrow{C} = c_0, c_1, \ldots, c_{P-1}$ as a long PN code, where P is the period of the long PN code. We take a segment from the long PN code to spread one signal bit. Assume the length of each PN segment is 1, that is, we use 1 chip to spread one signal bit. c_j represents one chip and c_j is either 1 or -1. We assume that bits x_i and x_j ($i \neq j$) are independent. The modulated signal $\overrightarrow{X}$ can be written as follows,

$$\overrightarrow{X} = (x_0\overrightarrow{C_0}, x_1\overrightarrow{C_1}, \ldots, x_{N-1}\overrightarrow{C}_{N-1}) \tag{11}$$

$$= (x_0c_0, \ldots, x_0c_{l-1}, x_1c_l, \ldots, x_1c_{2l-1}, \ldots, x_{N-1}c_{(N-1)l}, \ldots, x_{N-1}c_{Nl-1}) \tag{12}$$

Since x_i is independently and identically distributed, $P(x_ic_j = A) = 1/2$ and $P(x_ic_j = -A) = 1/2$, thus $E(x_ic_j) = 0$ and the standard deviation $\delta = A$. The following formula can be used to estimate the autocorrelation of a time series represented by $\overrightarrow{X}$,

$$r(\gamma) = 1/(N-\gamma)\sum_{i=0}^{N-1-\gamma}(a_{i_j} * a_{i_j} + \gamma) \tag{13}$$

where γ is the lag, $a_{i_j} = x_ic_{il+j}$ is the i^{th} item of $\overrightarrow{X}$, and $i \in [9, N-1], j \in [0, l-1]$.

The MSAC method reveals the presence of short PN code based DSSS watermarks by calculating $E(r^2(\gamma))$. $r^2(\gamma)$ is the square autocorrelation of spread signal $\overrightarrow{X}$ and a time-shifted $\overrightarrow{X}$ with lag γ. By calculating $E(r^2(\gamma))$, periodic peaks with a period of l will show up.

Theorem 1 shows there are no periodic peaks in our long PN code based watermarking technique under this MSAC detection method. The long PN code based DSSS watermarking technique is invisible for suspect *sender* and *receiver*. The detailed proof of Theorem 1 is in Appendix B of our technical report [14].

Theorem 1: The mean value of $E(r^2(\gamma))$ is

$$E(r^2(\gamma)) \approx \begin{matrix} A^4, \gamma = 0 \\ 0, \gamma \neq 0 \end{matrix} \tag{14}$$

According to Theorem 1, it is secret to use long PN code based DSSS watermarking technique to trace traffic flows since there is only one peak shown in the MSAC detection method at the lag $\gamma = 0$. Unlike using the short PN code based DSSS watermarking technique in [2], which reveals the self-similarity of embedded DSSS watermarks occurring at regular intervals, no periodic peaks show up for the long PN code based traceback. The traceback invisibility is preserved against MSAC analysis.

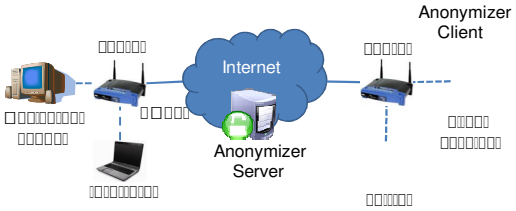
4 Evaluation

We conducted real-world experiments on Anonymizer to evaluate the performance of the long PN code based DSSS watermarking technique. In this section, we will first introduce the experiment setup. We will then present experimental results of detection rate and false positive rate and the capability of the new traceback approach on tracing

multiple flows. Finally, we demonstrate the long PN code based technique can defeat the MSAC based watermark detection.

4.1 Experiment Setup

Figure 5 illustrates the experiment setup. A web server *sender* running Windows 7 is located at a university campus. An off-campus computer *receiver* runs an Anonymizer client, which connects through an encrypted wireless network to the Anonymizer server. By setting up an encrypted VPN tunnel between the off-campus computer and Anonymizer server on the Internet, the off-campus computer can surf the web without exposing its real IP address. In order to determine if the off-campus computer is downloading a file from the web server, we use a computer as *interferer* to interfere with the outbound wireless traffic from the web server, and use another computer as *sniffer* to sniff the inbound wireless traffic to the receiver. The interferer and the sender are connected by a router, as are the sniffer and the receiver. The interferer and the sender share a link, so that interferer can interfere with the sender's traffic and modulate the outbound traffic with the long PN code based approach. This setup is a typical communication scene in an ad-hoc wireless network. In case of conducting network forensics on household wireless networks where the web server is wired into the Internet, law enforcement can interfere the traffic along the path from the web server to the client, for example, at an intermediate router.



an advantage of the long PN code based traceback. In the experiments, we first generate a long PN code of $2^{15}-1$ chips, and use the mask, {*0*, *0*, *0*, *0*, *0*, *0*, *1*, *0*, *0*, *0*, *1*, *1*, *0*, *0*, *0*}, to generate a long PN code with shift 20473. We then use segments of the shifted long PN code to spread the signal {1 -1 1 1 -1 1 -1}, as discussed in Section 3. The chip duration is fixed at 1 second.

We first examine the impact of the interfering CBR traffic rate (watermark amplitude) on detection rate. We change the CBR packet sending frequency from $1 packet/5ms$ to $1 packet/100ms$. In Figure 6, we can see that when the CBR traffic rate decreases (packet sending interval increases), the detection rate decreases. This is because slow interfering traffic incurs small watermark amplitude A in (5).

We then examine the impact of different long PN code lengths on detection rate. We used different long PN code segment lengths from 1 to 7 to spread a signal bit. Figure 7 shows that in general longer segment length achieves higher detection rate. This is the benefit of using spread spectrum spreading: we can use a long code to fight a noisy environment for better performance.

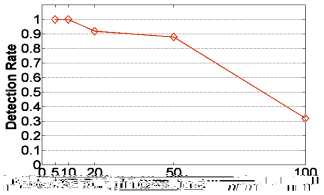


Figure 6: Detection Rate vs. Pkt. Sending Interval

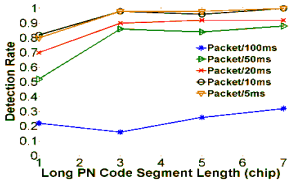


Figure 7: Detection Rate vs. Segment Length

We also examine the impact of chip durations on detection rate. We varied the chip duration from 0.3s to 1.5s. From Fig.8, we can see that under different chip durations, the detection rate has fluctuation but is consistently high overall.

4.3 False Positive Rate

Recall that the false positive rate $P_{F,n}$ for recognizing a n-bits original signal is $P_{F,n}=1/2^n$ [2]. In our experiments, we varied the signal length from 1 to 7. For each signal length we measured the false positive rates for the long PN code segments of different lengths from 2 to 7. The false positive rate for each signal length is calculated as the average of the probabilities of detecting the signal with different long PN code segment lengths. From Fig. 9, we can see that the false positive rate decreases with the increasing long PN code segment length. The theoretical curve matches the empirical curve very well.

4.4 Defeating MSAC Detection

In [2], the authors investigated the detection of watermarks generated by a short PN code, which is used to spread each signal bit. Through the *mean-square autocorrelation* (MSAC) analysis, periodic peaks show up due to self-similarity in the modulated traffic caused by homogeneous PN codes that are used in modulating a multiple-bit signal. Our strategy can defeat the MSAC analysis since we use different long PN code segments to spread different signal bits. Figure 10 shows the MSAC of a modulated flow. We can see there is no periodical peak any more. The authors also used detection rate P_D and false positive rate P_F as evaluation metrics for evaluating MSAC's capability to detect short PN code generated DSSS watermarks. When they try to detect traffic containing DSSS

watermarks, they need a high detection rate and a low false positive rate. Figure 11 shows Receiver Operating Characteristic (ROC) curve for our long PN code generated watermarks, which is a plot of P_D versus P_F. It can be observed that the false positive rate is as high (or low) as the detection rate. Therefore, it is hard to detect long PN code generated watermarks by the MSAC analysis.

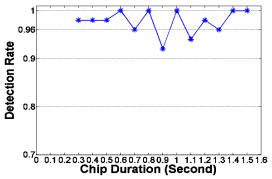


Figure 8: Detection Rate vs. Chip Duration

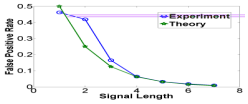


Figure 9: False Positive Rate

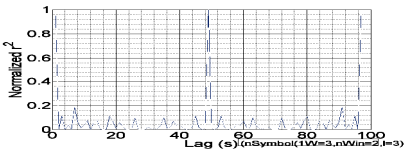


Figure 10: Estimation of MSAC

Figure 11: ROC

5 Conclusions

In this paper, we propose a long PN code based DSSS watermarking technique to trace suspect communication over encrypted (and open) wireless networks and anonymous communication networks on the Internet. This traceback technique has good invisibility. Since different segments of a long PN code are used to modulate different signal bits, this technique removes regular patterns and self similarity from the generated watermarks. Therefore, it can defeat mean-square autocorrelation (MSAC) based detection of watermarks generated by a short PN code, which is used to repeatedly modulate each signal bit.

Through a combination of analytical modeling and an extensive set of experiments over WLAN and Anonymizer, we demonstrated the effectiveness of the long PN code based DSSS watermarking technique. The long PN code based DSSS watermarking technique is a general one and can be used in other cyber crime scene investigations.

References

[1] Anonymizer, Inc., http://www.anonymizer.com/, 2010.

[2] Yu, W., X. Fu, S. Graham, D. Xuan, and W. Zhao. *Dsss-based flow marking technique for invisible traceback*. In Proceedings of the 2007 IEEE Symposium on Security and Privacy (S&P), pages 18-32, May 2007.

[3] Jia, W., F. Tso, Z. Ling, X. Fu, D. Xuan, and W. Yu. *Blind detection of spread spectrum flow watermarks*. In Proceedings of the 28th IEEE International Conference on Computer Communications (INFOCOM), Rio de Janeiro, Brazil, pages 2195-2203, April 2009.

[4] Zhu, Y., X. Fu, B. Graham, R. Bettati, and W. Zhao. *On flow correlation attacks and countermeasures in mix networks.* In Proceedings of Workshop on Privacy Enhancing Technologies (PET), pages 207-225, May 2004.

[5] Levine, B. N., M. K. Reiter, C. Wang, and M. Wright. T*iming attacks in low-latency mix-based systems*. In Proceedings of Financial Cryptography (FC), pages 251-265, February 2004.

[6] Murdoch, S. J., and G. Danezis. *Low-cost traffic analysis of tor.* In Proceedings of IEEE Security and Privacy Symposium (S&P), pages 183-195, May 2006.

[7] Fu, X., Y. Zhu, B. Graham, R. Bettati, and W. Zhao. *On flow marking attacks in wireless anonymous communication networks.* In Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS), pages 493-503, April 2005.

[8] Overlier, L., and P. Syverson. *Locating hidden servers.* In Proceedings of the IEEE Security and Privacy Symposium (S&P), pages 100-114, May 2006.

[9] Kiyavash, N., A. Houmansadr, and N. Borisov. *Multi-flow attacks against network flow watermarking schemes.* In Proceedings of the 17th USENIX Security Symposium, pages 307-320, July/August 2008.

[10] Peterson, W. W., and E. J. Weldon. *Error-Correcting Codes*. 2nd Edition. Cambridge, MA: The MIT Press, 1972.

[11] Lee, S. *Spread Spectrum CDMA: IS-95 and IS-2000 for RF Communications.* Chicago, IL: McGraw-Hill Professional, August 2002.

[12] Oppenheim, A. V., A. S. Willsky, and S. H. Nawab. *Signals and Systems*. 2nd ed. Upper