Anonymous Web Browsing through Predicted Pages

Shui Yu, Theerasak Thapngam, Hou In Tse
School of Information Technology,
Deakin University
Burwood, VIC 3125, Australia
Email: {syu, tthap, hit}@deakin.edu.au

Jilong Wang[1,2]
[1]Network Research Center, Tsinghua University.
National Lab for Information Science and Technology,
[2]Tsinghua University, Beijing 100084, P. R. China
Email: wjl@cernet.edu.cn

***Abstract*—Anonymous web browsing is an emerging hot topic with many potential applications for privacy and security. However, research on low latency anonymous communication, such as web browsing, is quite limited; one reason is the intolerable delay caused by the current dominant dummy packet padding strategy, as a result, it is hard to satisfy perfect anonymity and limited delay at the same time for web browsing. In this paper, we extend our previous proposal on using prefetched web pages as cover traffic to obtain perfect anonymity for anonymous web browsing, we further explore different aspects in this direction. Based on Shannon's perfect secrecy theory, we formally established a mathematical model for the problem, and defined a metric to measure the cost of achieving perfect anonymity. The experiments on a real world data set demonstrated that the proposed strategy can reduce delay more than ten times compared to the dummy packet padding methods, which confirmed the vast potentials of the proposed strategy.**

I. Introduction

The purpose of this paper is to present a novel strategy to achieve perfect anonymity in web browsing. Anonymous web browsing is demanded by Internet users for privacy and security reasons, yet web browsing with perfect anonymity has rarely been achieved using traditional methods, such as packet padding. According to the recent ACM survey [1], anonymous communication systems can often be classified into two general categories: high-latency systems and low-latency systems. High-latency anonymity systems are able to provide strong anonymity, but are typically only applicable for non-interactive applications that can tolerate delays of several hours or more, such as the mix networks [2] for email messages; On the other hand, low-latency anonymity systems often provide better performance and are intended for real-time applications, particularly web browsing. The examples for this category are the Tor system [3], the Crowds system [4]. Because of the strict time constraint, it poses a great challenge to achieve perfect anonymity in the low-latency systems, such as web browsing. The general goal of the attacks on anonymous communication is to identify pairwise entities in systems, rather than the content of communication, and the content is usually encrypted. For example, does user A communicate with user C? Does user D access web site E?

Data encryption is usually hired to provide security, however, data encryption for anonymous web browsing is vulnerable to traffic analysis. In general, every web page is different from the others. Differences include length of HTML text, number of web objects, number of packets for each web object and timing information of packet transportation. Each web site has its own distinctive feature which is a combination of all features of its web pages. In this paper we use *fingerprint* to represent the uniqueness of a web page. A web site may be accessed in an encrypted way, such as using SSL, however, encryption brings limited changes on the fingerprint [5]. As a result, attackers usually use traffic analysis techniques to break anonymity in communication [6] [7].

Packet padding techniques are normally employed on top of data encryption to protect anonymity in web browsing. Perfect anonymity in communication means that the anonymity cannot be breached in any situation. According to Shannon's perfect secrecy theory [8], perfect anonymity in communication is possible. Researchers currently deploy packet padding techniques to disguise fingerprints in communication. For example, hiding fingerprints for traffic sessions [9] [10]; Moreover, in order to hide timing information of connections, link padding is employed [11]. Currently, the dominant strategy of packet padding is using dummy packets as cover traffic. This strategy results in two major problems for communication: huge delay and extra bandwidth demand. Because of the strict delay constraints from web viewers, it is almost impossible to achieve perfect anonymity in web browsing using the dummy packet padding strategy.

In this paper, we propose a novel approach to address the aforementioned problem, which is an extension of our previous work [12]. In our previous work, we have shown the great potential of obtaining anonymity for web browsing using prefetched web pages, we explore further in this paper in various aspects. Our proposal comes from the fact that users generally access a number of web pages at one web site according to their own habits or interests. This has been confirmed by applications of web caching and web page prefetching technologies [13] [14]. Therefore, we can use the prefetched data to replace dummy packets for padding. We propose to disguise the fingerprints of web sites at the server side by injecting predicted web pages that users are going to download as cover traffic, rather than using dummy packets as cover traffic. From a long term viewpoint, this novel strategy wastes limited bandwidth and causes limited delay.

The contributions of this paper are listed as follows.

- We proposed to use predicted web pages to conduct packet padding for web browsing, rather than using dummy packets, which fundamentally addresses the prob-

978-1-4244-8865-0/10/$26.00 ©2010 IEEE

lems of extra delay and extra bandwidth demand of the traditional dummy packet padding methods.
- We established a simple mathematical model based on the perfect secrecy theory for the proposed strategy and defined a metric to measure the cost for web browsing with perfect anonymity.

The rest of this paper is structured as follows. Section II introduces the related work. We present the setting of the problem in Section III, and followed by the system modelling and analysis in Section IV. The preliminary performance evaluations are conducted in Section V. Finally, Section VI summarizes the paper and discusses future work.

II. Related Work

The HTTP protocol document [15] shows that when a client submits an HTTP request to an URL, the corresponding server will deliver the HTML text to the client, and the HTML text includes the references of the related objects, e.g. images, flashes. The objects will be downloaded to the client one after the other. Therefore, each web page has its own fingerprint in terms of number of web objects, packet arrival time intervals, and so on. Some web server may encrypt the content of packets, however, the fingerprint cannot be disguised by the encryption against traffic analysis.

A number of works have been done in terms of traffic analysis. Sun et al. tried to identify encrypted network traffic using the HTTP object number and size. Their investigation shows it is sufficient to identify a significant fraction of the world wide web sites quite reliably [5]. Following this direction, Wright, Monrose and Masson further confirmed that web sites can be identified with high probability even it is encrypted channel [16] . Hintz [10] suggested to add noise traffic (also named as *cover traffic* in some papers) to users which will change the fingerprints of the server, and transparent pictures are employed to add extra fake connections against fingerprint attacks.

Researchers also explored profiling attacks and proposed solutions. Timing attacks [17], [18] based on the fact that low-latency anonymous systems, such as onion routing, do not introduce any delays or significant changes on the timing patterns of an anonymous connection. The time intervals of the arrival packets of HTML text and HTTP objects usually similar for the target user and the adversary if they access the same web page, then it is easy for the adversary to figure out which web site the target user accessed from the list. Coull et al. evaluated the strength of the anonymization methodology in terms of preventing the assembly of behavioral profiles, and concluded that anonymization offers less privacy to web browsing traffic than what we expected [19] . Liberatore and Levine used a profiling method to infer the sources of encrypted HTTP connections [20] . They applied packet length and direction as attributes, and established a profile database for individual encrypted network traffic. Based on these information, they can infer the source of each individual encrypted network traffic. The match technique based on a similarity metric (Jaccard's coefficient) and a supervised learning technique (the naive Bayesian classifier). Their extensive experiments showed that the proposed method can identify the source with the accuracy up to 90%.

Wright, Coull and Monrose recently proposed a traffic morphing method to protect the anonymity of communication [21]. They transformed the intended web site (e.g. www.webmd.com) fingerprint to the fingerprint of another web site (e.g. www.espn.com). The transformation methods that they took include packet padding, packet splitting. Optimal techniques are employed to find the best cover web site (in terms of minimum cost for transformation) from a list. They tested their algorithm against the data set offered in Liberatore and Levine's work [20], and found that the proposed method can improve the anonymity of web site accessing and reduce overhead at the same time. Venkitasubramaniam, He and Tong [9] noticed the delay caused by adding dummy packets into communication channels, and proposed transmission schedules on relay nodes to maximize network throughput given a desired level of anonymity. Similar to the other works, this work is also based on the platform of dummy packet padding.

Web caching and prefetching are effective and efficient solutions for web browsing when bandwidth is a constraint. Award, Khan and Thuraisingham tried to predict user web surfing path using a hybrid model, which combines Markov model and Supporting Vector Machine [22]. The fusion of the two models complements each other's weakness and worked together well. Montgomery et al. found that the information of user's web browsing path offers important information for a transaction, and a successful deal usually followed by a number of same path accessing [23]. Teng, Chang and Chen argued that there must be an elaborate coordinating between client side caching and prefetching, and formulated a normalized profit function to evaluate the profit from caching an object [13]. The proposed function integrated a number of factors, such as object size, fetching cost, reference rate, invalidation cost, and invalidation frequency. Their event-driven simulations showed that the proposed method performed well.

III. Problem Setting

Alice accesses web sites via encrypted channels, and an adversary (Bob) focuses on identifying which web site Alice chooses from a list of possible web sites. We suppose Bob has the knowledge of all the web sites on the list. Bob also captures all the network traffic of Alice's computer. We assume there are n possible web sites that Alice accesses, $\{w_1, w_2, \ldots, w_n\}$. The a priori of $w_i (1 \leq i \leq n)$ is denoted as $p(w_i)(1 \leq i \leq n)$. For each web site $w_i (1 \leq i \leq n)$, we denoted its fingerprint as $\{p_i^1, p_i^2, \ldots, p_i^k\}$. For example, for a given web site w_i, we counted the number of packets for every web object, and saved them as $\{x_1, x_2, \ldots, x_k\}$. We unified this vector and obtained the distribution as $\{p_i^1, p_i^2, \ldots, p_i^k\}$, where $p_i^j = x_j \cdot (\sum_{m=1}^{k} x_m)^{-1}, 1 \leq j \leq k$. Bob monitored Alice's local network, and obtained a number of observations.

$$\tau = \{\tau_1, \tau_2, \tau_3, \ldots\} \quad (1)$$

Based on these observations and the Bayesian theorem, Bob is able to claim that Alice accesses web site w_i with the following probability.

$$p(w_i \mid \tau) = \frac{p(w_i) \cdot p(\tau \mid w_i)}{p(\tau)} \tag{2}$$

where $p(w_i)$, $p(\tau \mid w_i)$ and $p(\tau)$ are known to Bob, because Bob can actually access the n web sites individually to obtain these information.

On the other hand, the task for Alice is to decrease $p(w_i \mid \tau)$ to the minimum. As we know that data encryption itself cannot achieve the goal of anonymity, therefore, Alice has to further employ anonymization operations, such as packet padding, link padding to fight against Bob.

According to Shannon's perfect secrecy theory [8], an adversary cannot break the anonymity if the following equation holds.

$$p(w_i \mid \tau) = p(w_i) \tag{3}$$

Namely, the observation offers no information to the adversary. However, the cost for perfect anonymity is extremely expensive by injecting dummy packets according to the perfect secrecy theory. In order to measure the cost for perfect anonymity, we made a definition as follows.

Definition. *Cost Coefficient of Anonymity*. Let function $C(S)$ represent the cost function for a given network traffic S. For a given intended network traffic X, we inject a cover traffic Y to achieve the goal of anonymity, then the cost coefficient of anonymity is defined as

$$\beta = \frac{C(Y \mid X) + C(X)}{C(X)} \tag{4}$$

This metric will be used to indicate the cost efficiency for perfect anonymity operations in this paper.

IV. System Modeling and Analysis

In real applications, accessing patterns could be very complex. However, in this paper, we target on presenting the novelty, effectiveness of our proposed strategy as a new anonymization method, and demonstrating the great potential of the proposed strategy. Therefore, we confined our research space with the following conditions:

- We focused on perfect anonymity of network traffic sessions and ignored the link padding issue.
- We only discussed the cases on achieving anonymity by packet padding and excluded the packet splitting operations.
- We used the number of packets of web page objects as the fingerprint and focused on this kind of attacks in this context, and we did not discuss the timing traffic analysis attacks in this paper.

A typical anonymous web browsing system with data encryption (at the Internet channels) and packet padding (at the server side) is shown in Figure 1. As a client, Alice sends a HTTP request to a web server w_i via an encrypted channel. The web server also employes an encrypted channel

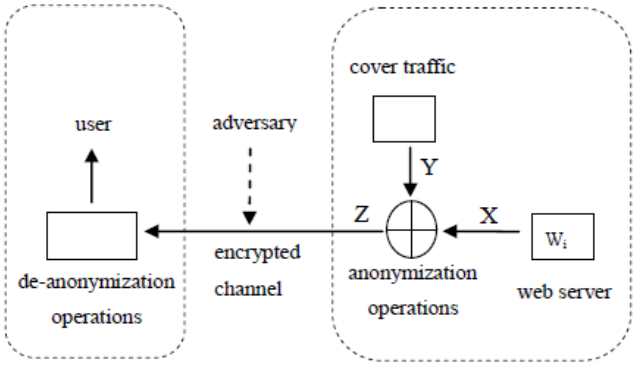

Fig. 1. A packet padding system for anonymous web browsing

to return the intended traffic $X = \{x_1, x_2, \ldots, x_k\}$, where $x_i (1 \leq i \leq k)$ represents the number of packets of web object i. Let $||X|| = \sum_{i=1}^{k} x_i$ denote the total number of packets of the intended traffic X. We then extracted the fingerprint of this session as $p = \{p_1, p_2, \ldots, p_k\}$ where $p_i = x_i \cdot (||X||)^{-1}, 1 \leq i \leq k$, and $\sum_{i=1}^{k} p_i = 1$. In order to make it anonymous to adversaries, we created cover traffic $Y = \{y_1, y_2, \ldots, y_k\}$ at the server side. Let $y_i (1 \leq i \leq k)$ denote the number of packets assigned to cover x_i, and let $||Y|| = \sum_{i=1}^{k} y_i$. Similar to the intended traffic X, the fingerprint of Y is $q = \{q_1, q_2, \ldots, q_k\}$. If $Z = \{z_1, z_2, \ldots, z_k\}$ represents the mixture of the intended traffic X and the cover traffic Y, then the fingerprint of Z is $r = \{r_1, r_2, \ldots, r_k\}$, and the total number of the mixed traffic is $||Z||$. The adversary's observation τ is the mixture of Z and other background traffic on the network.

In previous works, dummy packets are employed to work as the cover traffic Y; Once Z arrives at the client side, the dummy packet Y will be discarded; another solution is using transparent images as the cover traffic. However, in the proposed strategy the predicted web data is used as the cover traffic Y, the client decompose the received traffic, and the intended traffic X goes to the web browser, and the prefetched data Y is stored in the cache of the local computer, and Y may be used by the following requests. In this case, the client will fetch the expected web data from the cache, rather than download it again from the server. From a long term viewpoint, the bandwidth is not wasted and the average extra delay is limited in the proposed scheme.

In order to achieve perfect anonymity as described by Shannon [8], the following equation must hold.

$$r_1 = r_2 = \ldots = r_k \tag{5}$$

Furthermore, the following condition must also hold.

$$z_1 = z_2 = \ldots = z_k \tag{6}$$

This means that every traffic session is the same. As a result, Bob cannot obtain any information from his observation.

Let $x_{max} = max\{x_1, x_2, \ldots, x_k\}$, then the minimum cover traffic to achieve perfect anonymity is given as follows.

$$\begin{cases} y_1 = x_{max} - x_1 \\ y_2 = x_{max} - x_2 \\ \cdots \\ y_k = x_{max} - x_k \end{cases} \quad (7)$$

Let the cost function $C(\cdot)$ be the number of packets, then the cost coefficient for anonymity can be expressed as follows.

$$\beta = \frac{||Z||}{||X||} = \frac{\sum_{i=1}^{k}(x_i + y_i)}{\sum_{i=1}^{k} x_i} \quad (8)$$

Let β_d represent the cost coefficient for perfect anonymity using dummy packet padding strategy, then

$$\beta_d = \frac{||Z||}{||X||} = \frac{k \cdot x_{max}}{\sum_{i=1}^{k} x_i} \quad (9)$$

On the other hand, with the proposed strategy, the cover traffic is part of X in long term. The extra cost for the proposed mechanism is the part of the cover traffic that the client prefetches, but never accessed. We define the *missing rate* as the ratio of the unhitted prefetched data and the total prefetched data. Let $\eta (0 \leq \eta \leq 1)$ be the missing rate in the local cache, then the cost coefficient of perfect anonymity of the proposed strategy β_p is

$$\beta_p = \frac{\eta \cdot ||Y|| + ||X||}{||X||} = \frac{\sum_{i=1}^{k} \eta \cdot (x_{max} - x_i)}{\sum_{i=1}^{k} x_i} + 1 \quad (10)$$

Comparing equation (9) and equation (10), we obtained the following conclusion.

$$\beta_p \leq \beta_d \qquad (0 \leq \eta \leq 1) \quad (11)$$

The worst case for the proposed strategy is when $\eta = 1$ (all prefetched data is not used in the future), $\beta_p = \beta_d$ holds. In other words, our strategy is always not worse than dummy packet padding approaches.

V. Performance Evaluations

In order to confirm the advantages of the proposed strategy, we conducted two preliminary experiments using a real world data set [24], which is widely used by the community, such as in [20] and [21]. The data set includes the tcpdump files of 2000 web sites from February 10, 2006 to April 28, 2006. These have been sorted by popularity with all data encrypted. We took 30 continuous days from the most popular web site as the data set for the experiments in this paper, and treated each day as one session. We extracted the fingerprint (number of TCP packets for each web page object) of every session for the 30 days.

We first investigated the cost coefficient of perfect anonymity for the proposed strategy with different missing rate (namely, different prefetching accuracy). The results are shown in Figure 2.

We can see that when there is no missing ($\eta = 0$), the cost coefficient of perfect anonymity achieves the minimum, 1, which is the ideal case for users; When the missing rate is 0.5 ($\eta = 0.5$), the mean of the cost coefficient of perfect anonymity is around 7.59. Furthermore, the mean of the cost coefficient of perfect anonymity is 14.35 when all the prefetched data is missing ($\eta = 1$, equal the case of dummy packet padding strategy). This variation depends on the fingerprint distribution of the web pages. This preliminary experiment indicates that the cost for perfect anonymity using the dummy packet padding strategy is much higher (around 15 times more traffic volume and consequently, a 15 fold increase in delays) than using the proposed strategy. In other words, our strategy can reduce the delay up to 15 folds compare with the dummy packet padding method.

We are also interested in the relationship between the cost coefficient of perfect anonymity against the length of a session. We took 4 samples randomly from the data set over the

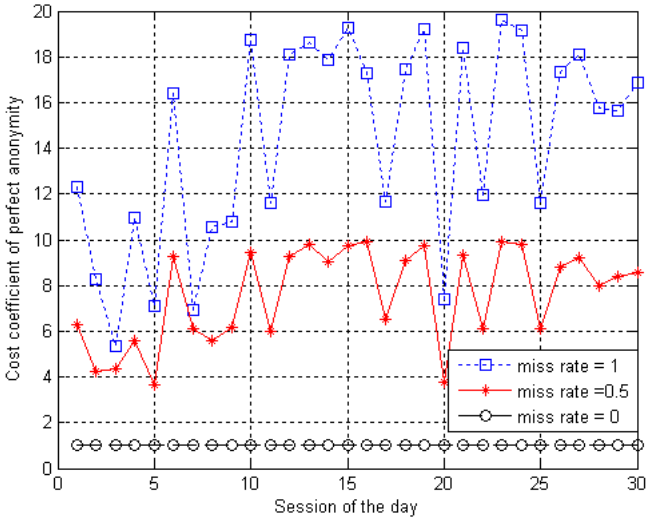

Fig. 2. Cost coefficient of perfect anonymity versus sessions with different missing rate.

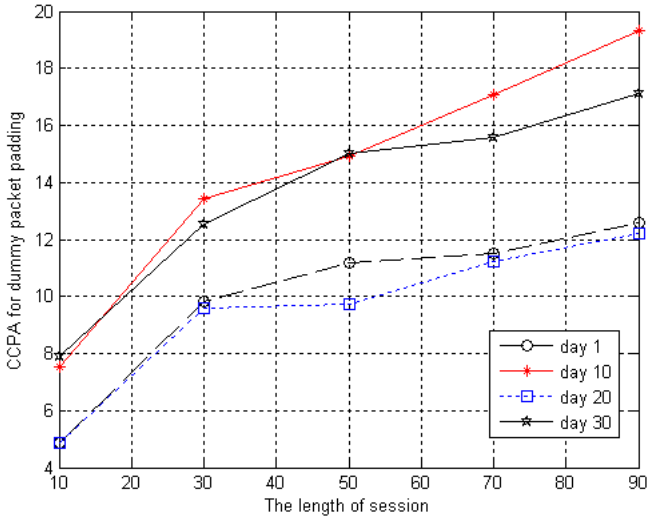

Fig. 3. Cost coefficient of perfect anonymity against length of session per day for the dummy packet padding strategy.

30 days period. We calculated the cost coefficient of perfect anonymity for the dummy packet padding strategy against the length of each session (we increased the length of sessions in the experiment). The results are shown in Figure 3. The cost coefficient of perfect anonymity of the dummy packet padding strategy was an increase function against the length of each session. Every change point indicated there were a bigger objects in terms of packet number in the past, namely, the change point depended on the distribution of larger objects. In other words, the longer the session length is, the higher cost for the dummy packet padding method in order to achieve perfect anonymity. However, this is not a problem for the proposed strategy.

VI. Summary and future work

In this paper, we proposed a novel strategy to achieve perfect anonymity in web browsing by using prefetched web pages as cover traffic, rather than using dummy packets as cover traffic. The proposed strategy makes web browsing with perfect anonymity much easy to achieve for Internet users, which is extremely hard to accomplish using the traditional dummy packet padding strategy. We have established a mathematical model for the problem based on Shannon's perfect secrecy theory, and our analysis showed that the proposed strategy was always equal to or better than the dummy packet padding strategy in terms of delay. Furthermore, the preliminary experiments confirmed our theoretical analysis, and demonstrated that the proposed strategy outperforms the traditional dummy packet padding method around 15 times in terms of delay and bandwidth cost.

The goal of this paper is to present a new method for anonymous web browsing. We have only revealed the huge advantages of the proposed method in this paper, and there are plenty of issues in this direction are not explored yet. We are currently exploring this new area through a wide investigation on the distribution of web site fingerprints and the theoretical relationship between fingerprint distribution and the cost for perfect anonymous web browsing.

In real applications on the Internet, it is extraordinary expensive to achieve perfect anonymity, therefore, alternative solutions are desperately expected. We list a few promising directions to share with our peers.

- Relative anonymization. In some cases, perfect anonymity may not necessary, as users only expect some level of anonymity for their web browsing; further, the adversary may not have the complete observation on the monitored users, therefore, an adaptive method may be introduced to further reduce the cost for anonymization. We believe the Game Theory can paly a great role in this direction, for example, finding the boundaries of anonymization cost against a given anonymity level.
- Packet dropping. It is interesting to investigate optimal strategies to reduce cost through the strategy of packet dropping. Moreover, it can change the outlook of fingerprint of a web page or web site against traffic analysis.
- Link padding has to be considered into the framework. The introduce of link padding will result huge cost, it poses a great challenge to achieve anonymity for web browsing.

References

[1] M. Edman and B. Yener, "On anonymity in an electronic society: A survey of anonymous communication systems," *ACM Computing Survey*, vol. 42, no. 1, 2009.

[2] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communication ACM*, vol. 24, no. 2, pp. 84–88, 1981.

[3] http://www.torproject.org.

[4] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for web transactions," *ACM Transaction on Information System Security*, vol. 1, no. 1, pp. 66–92, 1998.

[5] Q. Sun, D. R. Simon, Y.-M. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, "Statistical identification of encrypted web browsing traffic," in *IEEE Symposium on Security and Privacy*. Society Press, 2002.

[6] W. Yu, X. Fu, S. Graham, D. Xuan, and W. Zhao, "Dsss-based flow marking technique for invisible traceback," in *IEEE Symposium on Security and Privacy*, 2007, pp. 18–32.

[7] W. Jia, P. TSO, X. Fu, Z. Lin, D. Xuan, and W. Yu, "Blind detection of spread spectrum flow watermarks," in *INFOCOM'09: Proceedings of the 28th Conference on Computer Communications*, 2009.

[8] C. E. Shannon, "Communication theory of secrecy systems," *Journal of Bell System Technology*, vol. 28, pp. 656–715, 1949.

[9] P. Venkitasubramaniam, T. He, and L. Tong, "Anonymous networking amidst eavesdroppers," *IEEE Transactions on Information Theory*, vol. 54, no. 6, pp. 2770–2784, 2008.

[10] A. Hintz, "Fingerprinting websites using traffic analysis," in *Proceedings of the Workshop on Privacy Enhancing Technologies*, 2002.

[11] W. Wang, M. Motani, and V. Srinivasan, "Dependent link padding algorithms for low latency anonymity systems," in *ACM Conference on Computer and Communications Security*, 2008, pp. 323–332.

[12] S. Yu, T. Thapngam, S. Wei, and W. Zhou, "Efficient web browsing with perfect anonymity using page prefetching," in *Proceedings of the 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP10)*, 2010, pp. 1–12.

[13] W.-G. Teng and C.-Y. Chang, "Integrating web caching and web prefetching in client-side proxies," *IEEE Trans. Parallel Distrib. Syst.*, vol. 16, no. 5, pp. 444–455, 2005.

[14] M. Awad, L. Khan, and B. Thuraisingham, "Predicting www surfing using multiple evidence combination," *The VLDB Journal*, vol. 17, no. 3, pp. 401–417, 2008.

[15] http://www.w3.org/Protocols/rfc2616/rfc2616.html.

[16] C. V. Wright, F. Monrose, and G. M. Masson, "On inferring application protocol behaviors in encrypted network traffic," *Journal of Machine Learning Research*, vol. 7, pp. 2745–2769, 2006.

[17] V. Shmatikov and M.-H. Wang, "Timing analysis in low-latency mix networks: Attacks and defenses," in *ESORICS*, 2006, pp. 18–33.

[18] S. J. Murdoch and P. Zielinski, "Sampled traffic analysis by internet-exchange-level adversaries," in *Privacy Enhancing Technologies*, 2007, pp. 167–183.

[19] S. E. Coull, M. P. Collins, C. V. Wright, F. Monrose, and M. K. Reiter, "On web browsing privacy in anonymized netflows," in *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*. Berkeley, CA, USA: USENIX Association, 2007, pp. 1–14.

[20] M. Liberatore and B. N. Levine, "Inferring the source of encrypted http connections," in *CCS '06: Proceedings of the 13th ACM conference on Computer and communications security*. New York, NY, USA: ACM, 2006, pp. 255–263.

[21] C. V. Wright, S. E. Coull, and F. Monrose, "Traffic morphing: An efficient defense against statistical traffic analysis," in *Proceedings of the NDSS*, 2009.

[22] M. Awad, L. Khan, and B. M. Thuraisingham, "Predicting www surfing using multiple evidence combination," *VLDB Journal*, vol. 17, no. 3, pp. 401–417, 2008.

[23] A. L. Montgomery, S. Li, and K. Srinivasan, "Modeling online browsing and path analysis using clickstream data," *Marketing Science*, vol. 23, no. 4, pp. 579–595, 2004.

[24] http://traces.cs.umass.edu.