Packet Cloaking: Protecting Receiver Privacy Against Traffic Analysis

Reynold Cheng
Dept of Computing
Hong Kong Polytechnic U
Hung Hom, Hong Kong
csckcheng@comp.polyu.edu.hk

David K. Y. Yau
Dept of Computer Science
Purdue University
IN 47907, USA
Email: yau@cs.purdue.edu

Jianyun Fu
Dept of Computing
Hong Kong Polytechnic U
Hung Hom, Hong Kong
csjfu@comp.polyu.edu.hk

Abstract

With widespread use of the Internet, there is an increase in concern over users' privacy. In particular, an adversary may identify the receiver involved in a communication session by observing the packet traffic. We propose a new routing mechanism, which we call packet cloaking, to protect the privacy of a receiver. The main idea of packet cloaking is to transmit multiple copies of a sent packet to a selected group of k receivers, so that an adversary may only identify the true receiver with a probability of $\frac{1}{k}$. We present the system design to support packet cloaking. We also propose two metrics that measure the receiver privacy, based on evaluating the similarity between the sender/receiver traffic patterns. We have performed experimental evaluations to verify the effectiveness of our approach.

1 Introduction

The Internet is a prominent platform for wide area network communication. Numerous applications such as web-browsers, emails, and messaging software are being widely used. These applications are highly useful, but also pose important privacy concerns for their users. For example, a user surfing the net may not want to reveal the websites being visited. Unfortunately, user activities in the open Internet may be easily traced by eavesdroppers. For example, if it is observed that a user A is sending a sustained stream of data packets to another user B, over an extended period of time, then it is likely that A and B are engaged in a conversation with each other. It is therefore important to provide privacy while packets are being routed between users.

In this paper, we develop a routing mechanism to help avoid the receiver of a one-way communication from being identified through traffic observation. The mechanism, which we call *packet cloaking*, assumes that the network has excessive capacity, and introduces an amount of confounding traffic to increase the difficulty of tracing a communication. Specifically, given a sender and its intended recipient, a trusted entity called a *cloaking agent* finds $k-1$ other hosts in addition to the true recipient. The cloaking agent then sends each packet from the sender to these k recipients. Figure 1 shows a sender trying to send a packet to the receiver (circled in the figure) at IP address 127.8.3.2. The sender first sends the packet to the cloaking agent, specifying k to be 4. The cloaking agent accordingly generates four identical copies of the packet and sends them to the four receivers shown. Since the same packet is being sent to the four destinations at the same time, it is more difficult for an adversary to trace the packet to the true receiver. Moreover, if the four receivers are geographically far apart, it will be harder to localize the true receiver. Thus, it is possible to protect both the identify and the location of the receiver.

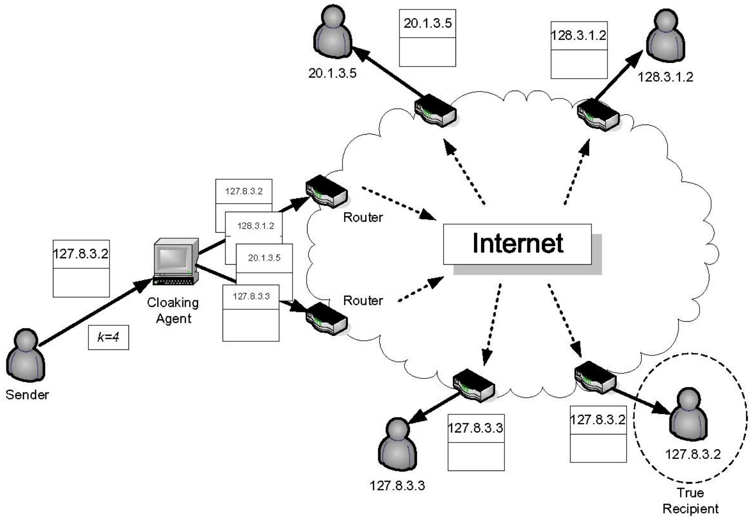

Figure 1. Packet Cloaking.

We now describe a system architecture to realize packet cloaking. The system encrypts the contents of a packet with a key known only to the packet's true receiver, thus providing content privacy [8] against unauthorized access. In selecting the receivers of a packet other than the true receiver, the cloaking agent will limit the candidate receivers to only those nodes that have agreed to participate in the cloaking system. To quantify the system's ability to provide privacy, we will investigate two new metrics based on k-anonymity [13] and time-series analysis [9, 15]. These metrics measure the linkability of the receiver with a given sender, and account for the receiver's privacy in terms of how easy the receiver's location (e.g., its subnet) can be identified. Additionally, we report ns-2 [6] simulation results to illustrate the cost of packet cloaking and its resilience to traffic analysis.

The rest of the paper is organized as follows. In Section 2, we review related work. In Section 3, we describe the detailed steps of packet cloaking. Section 4 develops two metrics for quantifying privacy. Experimental results are presented in Section 5. In Section 6 we conclude the paper.

2 Related Work

Information privacy has been a subject of active research in recent years. The foundation work in [10] presents definitions of terms related to privacy. We use the term "linkability" to refer to the ability of identifying the receiver-sender pair engaged in a communication. In [13], the notion of k-anonymity is defined, when the value of a data attribute is indistinguishable from $k - 1$ other values. Our method, which hides the true receiver among k receivers of the same packet, can be regarded as providing k-anonymity in packet routing. Recent work has also applied k-anonymity to provide location privacy [7, 3], where the location of a user can only be resolved to a region consisting of $k - 1$ other users. We similarly investigate the location privacy of the true receiver, relative to the locations of other possible receivers. (For example, whether the possible receivers are all in the same subnet.) Our work additionally investigates how the patterns of network traffic will impact on the privacy being provided. To that end, we introduce a new metric that combines both k-anonymity and time-series analysis.

We now summarize techniques for providing network privacy, which prevent packets from being traced in the network. In [2], an anonymous email system is proposed. The email system hides the identity of the receiver by introducing a number of "mix" nodes in the routing path. Mix nodes are machines that accept emails encrypted with their public keys, decrypt them, and send them to the next mix. The concept is further extended in onion routing [11, 4], where a packet is encrypted a number of times during the transmission. In [12], the notion of a *crowd* is proposed, in which a packet is routed through a number of random participating routers whose purpose is to confound, before the packet will be transmitted, by a probabilistic decision, to the intended final destination. Further to [2, 11, 4, 12], we quantify the impact of the pattern of network traffic on the privacy provided, using time-series analysis combined with k-anonymity. Recently, receiver location privacy using replicated packet transmissions is studied in [14], for the case of a wireless sensor network. The relationship between privacy and the characterization of network traffic is also not discussed in [14].

3 Packet Cloaking

In this paper, we assume a one-way communication between a sender and a receiver. The *cloaking agent*, being a central component in our privacy system, is used to generate packet traffics for the sender. In particular, for each communication request by a sender, the cloaking agent creates connections to k different recipients, one of which is the true recipient. Let us discuss this in more detail.

Before any transmission begins, a user needs to register with the cloaking agent for authorization to use the cloaking facilities. The user may also inform the cloaking agent about the conditions that it is willing to receive a cloaking packet. For example, a user may only want to help with packet cloaking when he is not using the computer, or if he is a good friend of the true recipient. These pieces of information are stored in the database of the cloaking agent. The agent then provides its public key to the new user. The user could also give his/her public key to the agent if he has not done so.

After registration, the sender may now send packets through the cloaking agent. Suppose that a sender, say S, wishes to send a packet to another user, say R. We assume that S has already signed the packet with its private key. The routing process comprises four steps (circled in Figure 2), as described below:

1. S encrypts the packet to be sent, using the public key provided by the cloaking agent, and signs the packet. The encryption and signature prevent the packet's contents from being read or modified while it is being sent to the cloaking agent. The packet is decrypted by the agent, and S's signature is verified. The cloaking agent then produces a list of k IP addresses, which it uses as the destinations of the sender's packet.

2. To obtain the k receivers, the cloaking agent consults its database. The database can be queried for a set of IP addresses corresponding to the nodes that are willing to receive a cloaking packet from S. The cloaking agent randomly selects $k - 1$ nodes from the set of candidates, in addition to the true receiver R. We call a receiver other than the true one a **cloaking receiver**.

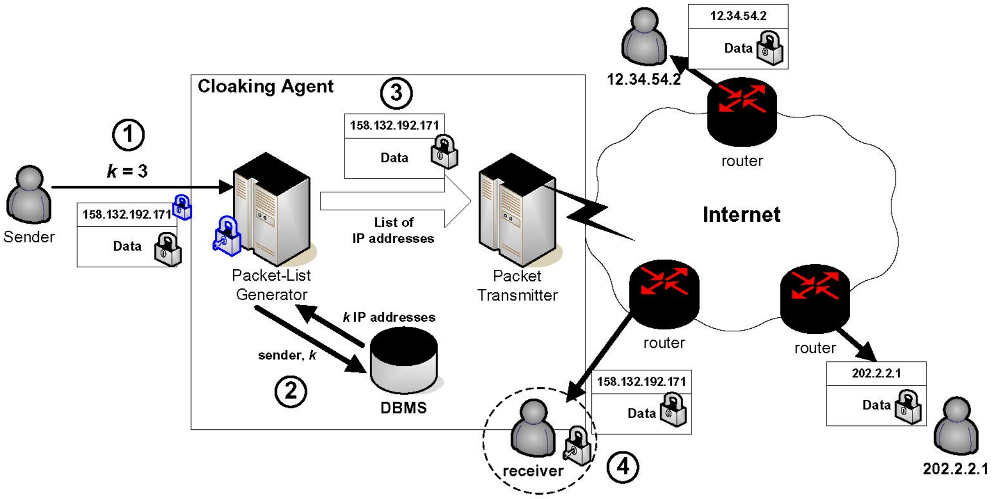

Figure 2. System Architecture for Packet Cloaking.

Notice that it may not be possible to find $k-1$ cloaking receivers. If this happens, the cloaking agent may ask the sender to use a smaller value of k. Also, this selection process only needs to be done once when S creates a connection with R; afterwards, the same list will be used throughout the connection.

3. The cloaking agent sends a packet to each of the k selected receivers. It first uses R's public key to encrypt the signed packet. Then, the packet transmitter of the cloaking agent produces k copies of the signed and encrypted packet, and send a packet to each of the receivers.

4. Upon receiving a packet sent by the cloaking agent, each of the k receivers attempts to decrypt the received packet. Since the packet is encrypted with R's public key, only R (who knows its own private key) can successfully read the packet. Each of the $k-1$ cloaking receivers will discard the packet when it fails to decrypt the packet.

In the process described above, the encryption/decryption are merely used to protect the *content privacy* [8] of the sender, and does not affect the effectiveness of protecting the receiver's location privacy – i.e., the contextual privacy [8] – against traffic analysis. Notice also that the sender's privacy may be better protected by not revealing its IP address in the packets being sent to the cloaking receivers, since these packets will be discarded anyway.

Another issue is that although only one packet transmitter is shown in Figure 2, it is possible to have more than one packet transmitter, which may not be necessarily in the same site. The use of more than one packet transmitter can help to distribute the load of generating packets, and also avoid the packet copies from overloading the cloaking agent's link to the Internet.

4 Correlation-based Privacy Metrics

To measure the effectiveness of packet cloaking against traffic analysis, we present two metrics quantifying the correlation between two traffic patterns. The first metric quantifies the linkability of a given sender with the true receiver, while the second one measures the location privacy of the receiver in terms of the probability that the receiver can be localized to its subnet.

Correlation Let $A = \{a_i | i = 1, 2, \dots, n\}$ and $B = \{b_i | i = 1, 2, \dots, n\}$ be two real-number sequences of length n. Let $\overline{A} = \sum_{i=1}^{n} \frac{a_i}{n}$ and $\overline{B} = \sum_{i=1}^{n} \frac{b_i}{n}$ be the respective means of A and B. Then the *correlation coefficient* between A and B, denoted by $c(A, B)$, is defined as follows [9, 15]:

$$c(A,B) = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{\sqrt{(\sum_{i=1}^{n}(a_i - \overline{A})^2)(\sum_{i=1}^{n}(b_i - \overline{B})^2)}} \qquad (1)$$

Equation 1 is a common way of comparing the similarity between two real-valued sequences. Note that $c(A, B) \in [0, 1]$, where a higher value of $c(A, B)$ indicates a higher degree of similarity between A and B.

Suppose an observer is able to monitor the traffic in the whole network for an extensive period of time. We measure the time in terms of time slots of equal length. Let the traffic observed at the output port of a sender for a fixed number m

of time slots be given by a real-valued sequence $S = \{s_i\}$, where s_i is the number of packets seen at time slot i. Let R_j (with $j = 1, 2, \ldots, N$) denote the jth node in a network of N nodes. Assume that the traffic observed at the input port of a node R_j for m time slots is given by the set $\{r_i^j\}$, where r_i^j is the number of packets observed on R_j at time slot i. If $H(j, t) = \{r_{i+t}^j\}$, where $t \in Z^+$, then $c(S, H(j, t))$ measures the similarity of the traffic pattern between the sender and the receiver R_j, right-shifted by some number t of time slots. The right shift is necessary because packets transmitted by the sender may experience a network delay before arriving at R_j. By considering this delay, the true correlation between the traffic patterns of the sender and the receiver can be measured. We now define two metrics based on correlation.

4.1 Sender-Receiver Linkability

Let the true receiver be R_r. Assume that the cloaking agent, on behalf of the sender, chooses a set K of k receivers, including R_r. We now define $P_r(t)$ as follows:

$$P_r(t) = \frac{c(S, H(r, t))}{\sum_{R_m \in K} c(S, H(m, t))} \qquad (2)$$

Essentially, $P_r(t)$ is the probability that R_r is identified through correlation with the sender's traffic, using a time delay t. Recall that in packet cloaking, the sender chooses k receivers each of which is potentially the true receiver. Hence, in the best case, the true receiver will be identified with a probability of $\frac{1}{k}$. However, if the observer can monitor the network for an extended period of time, the observer may be able to discover R_r by observing that the degree of traffic correlation between S and R_r is higher than that between S and any other receiver in K. Thus, Equation 2 captures the effectiveness of traffic observation in relating a sender with its receiver.

We can now define the *sender-receiver linkability*, denoted by $E_l(t)$, as follows:

$$E_l(t) = 1 - P_r(t) \qquad (3)$$

where E_l, ranging from 0 to 1, measures the linkability of the sender and its receiver. For example, if P_r is small, R_r has a lower probability of being identified as the true receiver. Correspondingly, E_l will be large.

4.2 Receiver's Location Privacy

The second metric evaluates the probability that the true receiver can be localized to its subnet. Let Y be the set of all hosts in the network. Further, let X be the set of hosts in the minimal subnet that includes the true receiver R_r. We now define $P_g(t)$ as follows:

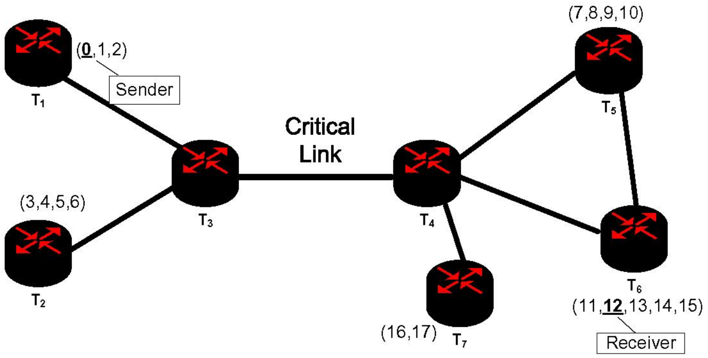

Figure 3. Network Topology for Experiments.

$$P_g(t) = \frac{\sum_{R_j \in X/\{S\}} c(S, H(j, t))}{\sum_{R_i \in Y/\{S\}} c(S, H(i, t))} \qquad (4)$$

In Equation 4, the numerator represents the sum of correlation values between S and each receiver residing in the same subnet as R_r. The denominator is the sum of correlation values for all the hosts in the network other than the sender. Intuitively, $P_g(t)$ is the probability that traffic patterns similar to the outgoing traffic from S can be observed in the same subnet of R_r. If P_g is high, a large volume of traffic similar to that received by R_r appears in R_r's subnet. If the IP address of the subnet can be associated with a physical location, then the observer can closely estimate R_r's location, even if the address of R_r is not known. Thus, P_g measures the probability that the receiver's location in a subnet is revealed.

We now define the *receiver's location privacy*, denoted by $E_g(t)$, as follows:

$$E_g(t) = 1 - P_g(t) \qquad (5)$$

where E_g, ranging from 0 to 1, measures the location privacy of a true receiver. A larger value of E_g represents a better protection for the receiver's location privacy.

5 Experimental Results

We have performed experiments to evaluate the proposed method of packet cloaking. We first describe the simulation setup. Then, we describe the detailed results.

5.1 Experimental Setup

We use BRITE [1] to generate the topology shown in Figure 3. There are seven routers ($T_1, T_2, \ldots, T_7$) and 18 hosts (indicated by the IDs shown under each router). Hosts that are connected to the same router form a subnet. For instance, R_{16} and R_{17} belong to the same subnet as they are connected to T_7. We have identified the link between routers R_3 and R_4 as the *critical link*, in the sense that most packets are routed through this link. The simulations are done in ns-2 using the generated topology. The bandwidth of each link between the routers is 100 Mb/s and the link's propagation delay is 30 ms. The bandwidth of a link between a router and a host is 10 Mb/s and the propagation delay is 3 ms.

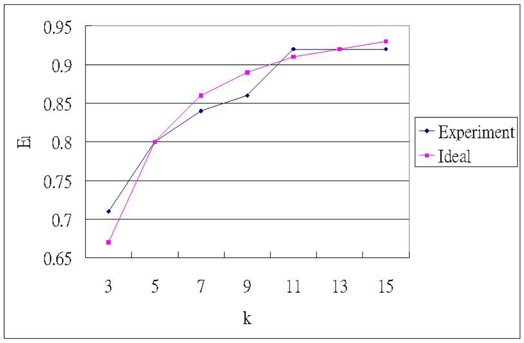

Figure 4. E_l **vs.** k**.**

We use host R_0 as the sender, and use host R_{12} as the intended receiver of R_0's packets. We assume that all the hosts are willing to participate in packet cloaking, so that the cloaking agent can choose among all the hosts in the generated topology as the cloaking receivers. Each packet has a size of 1,500 bytes, and is sent using UDP. By default, packets are generated using a Poisson process at an average rate of 2 Mb/s. We also use nine other active connections to simulate the background traffic in the Internet. To measure the correlation between the sender's traffic and the receiver's traffic to obtain the privacy measure in Equation 3, we use the delay of the first packet from the sender to the receiver as the time shift value. This is reasonable in our simulation, because a receiver gets no traffic until it receives the first packet from the sender. In practice, a more accurate correlation value may be computed by using more expensive analysis (e.g., finding the average propagation delay over a fixed period of time). In all our experiments, each reported data point is the average of 10 runs.

5.2 Results

5.2.1 Sender-Receiver Linkability

We first examine the effect of k on the sender-receiver linkability in terms of E_l. Figure 4 shows that E_l increases with the value of k. This is because when k increases, the new participants will receive packets from the sender, thereby increasing their correlation values (Equation 1). As a result, the probability of the true receiver being identified, P_r, decreases (Equation 2), causing a drop in the value of E_l (Equation 3). This shows that the sender-receiver privacy can be improved with a larger number of participants in the packet cloaking. In the same figure, we also compare our protocol's performance with an "ideal" graph, obtained by assuming that there is no other traffic except that generated by the cloaking agent. Thus, all the participants in the cloaking system will have the same traffic correlation value with the sender, and the true receiver is always identified with a probability of $P_r = \frac{1}{k}$. In this case, E_l becomes $1 - \frac{1}{k}$ or $\frac{k-1}{k}$. The performance of our protocol, considering the similarity between the sender and receiver traffic, is close to this curve. The small difference is due to the presence of the cross traffic used in the experiments.

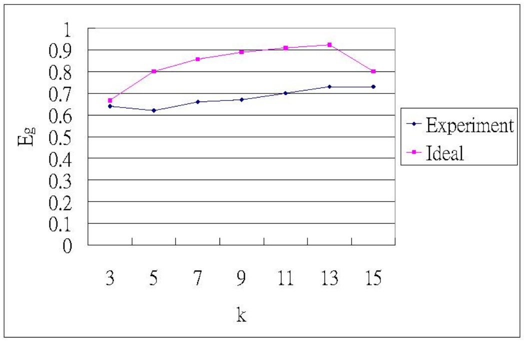

Figure 5. E_g **vs.** k**.**

5.2.2 Location Privacy of the True Receiver

Next, we investigate how the receiver's location privacy, measured in terms of E_g, is affected by changing the number of participants, k. Figure 5 shows the results over a wide range of k. We see that E_g is reasonably high (at least 0.6). The reason is that packet cloaking attempts to choose receivers in subnets different from the true receiver's subnet. As a result, packet traffic similar to that for the true receiver R_r, as generated by the cloaking agent, will have a small chance of reaching another receiver in R_r's subnet. Since the correlation values of the sender and the cloaking receivers are high, P_g is low (Equation 4) and E_g is high (Equation 5).

We also investigate the scenario when no traffic other than that generated by the cloaking agent is active in the network. We assume that the cloaking agent is lucky enough to always choose hosts not in the same subnet as R_r. Then, the nominator of Equation 4 only contains the correlation value of the true recipient, yielding a minimal value of P_g and correspondingly a maximal value of E_g. This "ideal" situation is shown in Figure 5. We can see that E_g increases with k until $k = 13$. This is because all the k cloaking receivers have the same correlation with the sender, while other nonparticipants have a zero correlation. Thus, Equation 4 reduces to $\frac{1}{k}$. The corresponding value of E_g becomes $1 - \frac{1}{k}$ which increases with k (Equation 5). When $k \geq 13$, E_g drops, since all the hosts in the subnets other than R_r's have been chosen as the cloaking receivers, thus forcing the hosts in the same subnet as the true receiver R_{12} (i.e., R_{11}, R_{13}, R_{14} and R_{15}) to be chosen as cloaking receivers (c.f. Figure 3). Hence the numerator of Equation 5 increases and E_g decreases. Note that the difference between the ideal case and the experiments is due to (1) the presence of cross traffic used in the experiments, and (2) the fact that hosts in the same subnet as R_r could be chosen.

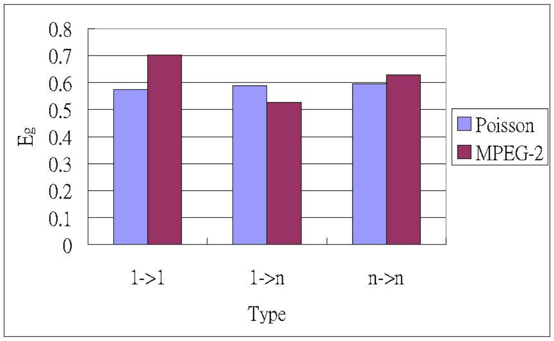

Figure 6. Poisson and MPEG traffic.

5.2.3 Effect of Traffic Types

This experiment compares two types of packet traffic: Poisson and a model of MPEG-2 [5]. The MPEG-2 traffic has a bit rate between 1Mb/s and 10Mb/s. The utilization of the critical link is set to be 20%. Figure 6 compares the performance for the two traffic types in three scenarios, namely: 1-to-1, 1-to-n, and n-to-n. Each of these cases depicts the correlation between the outgoing traffic from the sender (R_0) and the incoming traffic into the receiver (R_{12}). For example, a 1-to-1 connection means that R_0 and R_{12} have a single connection, while n-to-n means that both of the hosts have more than one connection. We can make two observations. First, although Poisson traffic is less regular than the MPEG-2 traffic, the privacy achieved by packet cloaking is insensitive to this difference. Second, the value of E_g is generally stable among the three different traffic scenarios. These results illustrate that packet cloaking can be applied under different scenarios, without significant degradation in the achieved privacy.

6 Conclusions

Protecting privacy in routing is important in Internet applications, when the network is subject to packet tracing or traffic observations. To provide privacy, we proposed a packet cloaking approach based on sending identical packets from a sender to a number of cloaking receivers, with permission by both the sender and the receivers. We presented the architecture of the proposed system. We also developed two privacy metrics to validate our methods.

Acknowledgments

The work described in this paper was supported by the Internal Competitive Research Grants (A-PH09 and A-PH39) of the Hong Kong Polytechnic University. We would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

[1] A. Medina et al. BRITE: An approach to universal topology generation. In *Proc. MASCOTS*, 2001.

[2] D. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–88, Feb 1981.

[3] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar. Preserving user location privacy in mobile data management infrastructures. In *Proc. 6th Workshop PET*, 2006.

[4] R. Dingledine, N. Mathewson, and P. Syverson. The second-generation onion router. In *USENIX*, 2004.

[5] Eldon Mellaney et al. Study of MPEG-2 video traffic in a multimedia lan/atm internetwork system. *IEEE Tran. Circuits and Systems for Video Technology*, 7(4), 1997.

[6] K. Fall and K. Varadhan. The ns manual, 2006.

[7] M. Gruteser and D. Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proc. 1st Intl. Conf. on Mobile Systems, Applications, and Services*, 2003.

[8] P. Kamat et al. Enhancing source-location privacy in sensor network routing. In *Proc. 25th IEEE ICDCS*, 2005.

[9] P. Zhang et al. Correlation analysis of spatial time series datasets: A filter-and-refine approach. In *PAKDD*.

[10] A. Pfitzmann and M. Hansen. Anonymity, unobservability, psuedonymity, and identity management - a proposal for terminology, Sept. 2004.

[11] M. Reed, P. Syverson, and D. Goldschlag. Anonymous connections and onion routing. *Journal on Selected Areas in Communications*, 16(4):482–294, May 1998.

[12] M. Reiter and A. Rubin. Crowds: Anonymity for web transactions. *TISSEC*, 1(1):66–92, June 1998.

[13] L. Sweeney. k-anonymity: a model for protecting privacy. *Intl. Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.

[14] Y. Jian et al. Protecting receiver location privacy in wireless sensor networks. In *Proc. INFOCOM*, 2007.

[15] Y. Sakurai et al. Braid: Stream mining through group lag correlations. In *Proc. SIGMOD*, 2005.