Priv-Code: Preserving Privacy Against Traffic Analysis through Network Coding for Multihop Wireless Networks

Zhiguo Wan*, Kai Xing‡, Yunhao Liu*†
*MOE Key Lab for Information System Security, School of Software,
Tsinghua National Lab for Information Science and Technology, Tsinghua University, wanzhiguo@tsinghua.edu.cn
†Department of Computer Science and Engineering, HKUST, liu@cse.ust.hk
‡School of Computer Science and Technology
University of Science and Technology, China, kxing@ustc.edu.cn

***Abstract*—Traffic analysis presents a serious threat to wireless network privacy due to the open nature of wireless medium. Traditional solutions are mainly based on the mix mechanism proposed by David Chaum, but the main drawback is its low network performance due to mixing and cryptographic operations. We propose a novel privacy preserving scheme based on network coding called Priv-Code to counter against traffic analysis attacks for wireless communications. Priv-Code is able to provide strong privacy protection for wireless networks as the mix system because of its intrinsic mixing feature, and moreover, it can achieve better network performance owing to the advantage of network coding. We first construct a hypergraph-based network coding model for wireless networks, under which we formalize an optimization problem whose objective function is to make each node have identical transmission rate. Then we provide a decentralized algorithm for this optimization problem. After that we develop an information theoretic metric for privacy measurement using entropy, and based on this metric we demonstrate that Priv-Code achieves stronger privacy protection than the mix system while achieving better network performance.**

I. Introduction

Traffic analysis is a powerful tool to deduce information from communication patterns, no matter whether the messages are encrypted or not. Numerous traffic analysis attacks have been successfully applied to various communication networks, including both military and civilian systems. Due to the open nature of wireless medium, wireless networks are specially vulnerable to traffic analysis attacks. Wireless telegraphy and radio network are two typical examples attacked by traffic analysis.

To fight against traffic analysis attacks, David Chaum [1] proposed the idea of mix to hide correspondence between message senders and receivers, and hence to protect communication privacy. Specifically, a Mix accepts a number of fixed-length messages from sources, performs cryptographic transformations on the messages, and then forwards the messages to the next destination in an order unpredictable from the order of inputs. An obvious feature of the mix-based system is that messages are cached and reordered at each Mix before being sent out, and the content of each message is not changed after the mixing process. Following the idea of Chaum, a number of anonymous communication systems have been proposed, including Crowds [2] and Tor [3] for web browsing, Mixmaster [4], and Mixminion [5] designed for email privacy.

Although the mix-based mechanism can be applied to protect privacy, its deficiencies are obvious. First of all, each Mix has to cache enough messages before sending them in a random order, which introduces unpredictable delay into the system. As a result, the mix-based system lacks ability to support realtime network traffic or guarantee certain quality of service. Next, the mix-based system requires each Mix to perform decryption and re-encryption, normally public key based operations, for each message. This leads to too much computation cost for Mixes, and also increases message transmission delay. Last but not least, efforts in this line of research have been focused on privacy preservation using cryptographic primitives, without considering network performance. Therefore it is imperative to design an anonymous communication system with the performance concern in the privacy preserving design, as the increasing popularity of resource-constrained wireless devices.

In this paper, we tackle the privacy-preserving communication system design problem based on the network coding technique. The concept of network coding was first proposed by Ahlswede et al. in 2000 [6], and has been viewed as a promising technology for improving network performance and enhancing network reliability. The network coding mechanism works differently from traditional routers or Mixes in that messages are coded at intermediate nodes, in contrast to message reordering in mix-based systems. Although network coding is proposed as a tool to improve network performance and reliability, it has the intrinsic mixing feature like the Mix in mix-based systems. Intuitively, Mixes in mix-based systems can be modified to encode messages instead of simple reordering for better performance and reliability.

However, there has been very limited research on employing network coding to counter against traffic analysis, and most work on traffic analysis focuses on mix-based systems. Until recently, the potential of network coding on resilience to traffic analysis has been noted by Fan et al. [7]. They analyzed privacy enhancement with network coding in case of traffic correlation attacks, while formal treatment of privacy improvement due to network coding scheduling is not given in their paper. And their scheme is designed for wired networks, without considering the broadcast nature of wireless medium.

In this paper, we propose Priv-Code, a network coding-based scheme to preserve privacy against traffic analysis for wireless communications. This scheme formalizes the privacy preserving problem as an optimization problem under a hypergraph-based network coding model for wireless networks. It provides a decentralized algorithm for this optimization problem. We develop an information theoretic metric for privacy measurement using entropy, and based on this metric we demonstrate that Priv-Code achieves stronger privacy protection than the mix system while achieving better network performance. The contributions of our work can be summarized as follows:

- We define a network model using the directed hypergraph similar to [8] for network coding scheduling in multihop wireless networks. The model not only captures the broadcast nature of wireless networks, but also considers the lossy characteristic and the MAC interference of wireless medium.
- We formalize the privacy-preserving network coding scheduling problem as an optimization problem, and provide a decentralized algorithm based on decomposition techniques.
- An information theoretic metric for quantitative privacy measurement based on information entropy is proposed in this paper. It provides a general way to evaluate privacy protection strength of various mechanisms with regard to traffic flow information.
- We implement the algorithm and conduct experiments for different network parameters to evaluate its privacy protection and performance. The proposed scheme not only provides strong privacy protection but also has good network performance.

The rest of the paper is organized as follows. In the next section we review related work on the mix system and network coding. Then our scheme on exploiting network coding to achieve privacy is described in detail in Section III. In Section IV we analyze and discuss issues of the proposed scheme on privacy protection. Details on the simulation to evaluate the proposed scheme are provided in Section V. Section VI summarizes and concludes the paper.

II. Related work

A. *The Mix-based Systems*

Since Chaum's mix system was proposed, many similar designs have been introduced in the literature [2], [3], [4], [5], [9]. In these systems, a *mix-net* is formed by a set of mix nodes, and messages are mixed when they traverse the mix-net to achieve anonymity. Generally, these systems can be grouped into real-time systems and non-realtime systems.

Crowds [2] and Tor [3] belong to the real-time mix systems. In Crowds, a user joins a "Crowd" as a "jondo", and each jondo acts as a proxy passing web requests to a random Crowd member or the web server according to a given probability. Thus, Crowds is able to preserve anonymity against collaborating members, but only receiver anonymity is provided in case of local eavesdropping. Cypherpunk, Mixmaster [4], and Mixminion [5] are non-realtime systems being used as anonymous re-mailers. Cypherpunk is the first widely implemented mix-like system, without features like the message padding or message pools, which makes it the "Type I" anonymous remailer. Mixmaster, the "Type II" anonymous remailer, fixed these problems in Cypherpunk, while the "Type III" remailer Mixminion further enhance privacy by making reply and forward messages indistinguishable. As analyzed, the main drawbacks of the mix-based systems are its low network performance and high computation cost.

Information theoretic treatment on privacy metric in mix-based systems is studied in [10]. The anonymity is measured by information entropy of the mix-based systems, which is able to accurately characterize privacy protection strength. The anonymity entropy measures the uncertainty that the attacker on identifying a sender or receiver. Let the probability distribution of a user u being the sender (or the receiver) of a message M be p_u, where $\sum\limits_u p_u = 1$, then the anonymity metric can be computed as $H(M) = -\sum\limits_u p_u \cdot \log p_u$. This entropy represents the anonymity status of the message M, and the number of bits of additional information the attacker needs to identify the message sender. In this paper, we use a similar information theoretic method to analyze privacy protection achieved for wireless networks with our scheme.

B. *Wireless Network Coding*

The network coding technique [6] was originally proposed as a solution to improve network performance and soon received extensive attention in the networking community. This technique has been extensively studied in wireless environments to fully exploit the broadcast nature of wireless medium.

Katti et al. [11] have implemented a real network coding system COPE that *XOR* packets in the wireless network. The performance improvement of COPE can be up to 70% for wireless mesh network, and COPE even has 3-4 times of throughput gain on the testbed.

Research on wireless network coding has studied network coding scheduling for different purposes or applications, ranging from broadcast, multicast, reliability, energy efficiency, to maximizing throughput.

Security and privacy in network coding have also been important research directions. There have been a lot of research on security issues in network coding, e.g., [12], [13], [14], [15], [16], [17], [18]. However, the privacy issue of network coding has been largely ignored by the research community.

Until recently, a scheme proposed by Fan et al. [7] is the only one to enhance privacy by network coding. They employed network coding to counter against traffic analysis attacks in wired networks, and use homomorphic encryption to protect code coefficients of messages. Although they analyzed privacy enhancement due to network coding in case of packet size correlation, time correlation and content correlation, formal treatment of privacy with regard to *traffic flows* is not given in their paper. That its, they did not answer the question on how to schedule network flows, the central part of network coding, in order to improve privacy. Also the proposed solution is designed only for wired networks, and cannot be used in wireless networks. Furthermore, the advantages of network coding on performance improvement and energy saving are not fully exploited for either wired networks or wireless networks.

III. The Proposed Scheme: Priv-Code

The proposed scheme Priv-Code considers concurrent unicast sessions in multi-hop wireless networks, and employs intra-session network coding for data communication, i.e., only packets from the same session can be encoded together. The main goal of Priv-Code is to achieve strong privacy against traffic analysis with proper network coding scheduling. The intuitive behind Priv-Code is that one can make all nodes in the network transmit traffic with the same *traffic pattern*, then the attacker is not able to distinguish traffic senders or receivers. If all nodes in the network transmit data with the same data rate, and all transmission flows look no different from each other for various traffic analysis attacks, then the attacker cannot obtain any information on senders or receivers at all. This can be done by using specially designed network coding scheduling which schedules an end-to-end unicast session over multiple paths. It will not only significantly enhances resistance against traffic analysis, but also improves network performance due to benefits of network coding.

In this section, we first introduce assumptions and threat model in our scheme, then we describe our system model which captures the broadcast nature and lossy characteristic of wireless medium. Next, we propose an optimization framework to find the optimal network coding scheduling for enhanced privacy. After that, we provide a decentralized algorithm for this optimization framework, which is specially designed for distributed multihop wireless networks.

A. Assumptions and Threat Model

In this paper, we assume a multi-hop wireless network, in which each node is equipped with only one antenna. All nodes in the network have identical transmission range as well as interference range. We also assume that the wireless medium is lossy. An anonymous routing protocol for multi-hop wireless networks providing anonymity and unobservability is implemented as [19]. Hence routes can be securely and anonymously established from a source to some destinations, and an outside attacker cannot access the packet header to know the packet type, or source/destination address. In order to avoid packet size correlation, all packets are of the same size.

We assume existence of a global adversary who can passively monitor the whole network. He can continuously observe the entire wireless network, and hence obtain traffic flow information including node transmission rates, inter-packet intervals etc. The attacker can make use of existing traffic analysis techniques used in [20] and [21]. However, he cannot decrypt any encrypted packet with brute force attack. His goal is to deduce who is the sender or receiver of a message from network traffic information.

B. The System Model

The data flow of a session is divided into generations, and packets from the same generation can be encoded. The number of packets n in a generation can be configured to suit the application. At each node, random linear network coding is used to process the packets, and the encoded packets are transmitted to the destination via multiple paths. How to establish multiple paths will be described in the next section.

The network is modeled as a directed hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{A})$ as in [8], where $\mathcal{N}$ is the set of nodes and $\mathcal{A}$ is the set of hyperarcs. A hyperarc (i, J) is formed by a start node i and a set of end nodes J, which is a non-empty subset of $\mathcal{N}$. Each hyperarc (i, J) represents a broadcast link from node i to nodes in J. As we assume all nodes have identical transmission range, each node i has a unique end node set J. This definition captures the broadcast nature of wireless medium. The hypergraph is degraded into a conventional graph model when J contains only one node. A set of unicast sessions $\mathcal{U} = \{u_1, ..., u_{|\mathcal{U}|}\}$ is transmitted through the network. Let S^k and T^k $(k = 1, 2, ..., |\mathcal{U}|)$ be the source and receiver of the unicast session k, and r^k denote the flow rate of session k.

For a unicast session k where the source S^k wants to send data with rate r^k to T^k, by the flow conservation condition, we have:

$$\sum_{j \in J} f^k_{iJj} - \sum_{j \in \mathcal{N}} \sum_{\{I|(j,I) \in \mathcal{A}, i \in I\}} f^k_{jIi} = \delta^k_i, \forall i \in \mathcal{N}, \qquad (1)$$

where

$$\delta^k_i = \begin{cases} r^k, & \text{if } i = S^k \\ -r^k, & \text{if } i = T^k \\ 0, & \text{otherwise} \end{cases}$$

and f^k_{iJj} is the flow rate over hyperarc (i, J) intended to node $j \in J$. For session k, the equation represents the flow conservation constraint that the source node's *net* transmission rate is r^k, the destination node's *net* transmission rate is $-r^k$, and any intermediate node's *net* transmission rate is 0.

Under the hypergraph model, we further set up the broadcast MAC model to characterize the interference in wireless networks, and the coding model how packets are coded with network coding.

We use the broadcast MAC model of Zhang and Li [22], which extends the unicast MAC model to obtain a necessary condition for feasible broadcast schedules. In this model, the transmission range and the interference range are considered to be the same, and the reception probability beyond this range can be ignored.

Specifically, the wireless network is modeled as an ideal time-slotted broadcast MAC where competing transmitters can optimally multiplex the channel without any collisions. For a unicast session k, let $B_i^k[t]$ (0 or 1) be the decision variable indicating whether node i is transmitting in slot t, and $I(i)$ be the set of all transmitters within i's range (including i). Under the hypergraph model, $I(i)$ is equal to node i's end node set J. Then a schedule is collision free iff:

$$\sum_k B_i^k[t] + \sum_k \sum_{j \in I(i)} B_j^k[t] \leqslant 1, \forall i \in \mathcal{N} \backslash S^k. \tag{2}$$

This equation indicates that any receiver i allows the broadcast transmission from at most one transmitter within its range at each time slot. Denote T as the period of a schedule, and b_i^k as the rate at which node i broadcasts packets to its downstream nodes, then we have:

$$b_i^k = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} B_i^k[t]. \tag{3}$$

Apply (3) to (2), we can have:

$$\sum_k b_i^k + \sum_k \sum_{j \in I(i)} b_j^k \leqslant C, \forall i \in \mathcal{N} \backslash S^k. \tag{4}$$

where $C = \frac{1}{T}$ is the MAC layer capacity, equaling to the maximal broadcast rate of a node when no interference presents.

Meanwhile, it is noted that a constraint on capacity for the broadcast link (i, J) should be satisfied. Since we assume lossy wireless links in our scheme, the reception probability of the link on the hyperarc (i, J) to node j is p_{iJj}. Then we have the following link capacity constraint:

$$b_i^k p_{iJj} \geqslant f_{iJj}^k. \tag{5}$$

C. Scheduling Optimization for Privacy

Before presenting the scheduling optimization algorithm for privacy, we first show how to make a traffic flow indistinguishable from another flow. The question is what kind of traffic pattern should the data flows take so as to make them indistinguishable. The traffic pattern of a flow is determined by packet arrival rate, inter-packet intervals and arrival time distribution. If message arrivals of traffic flows are Poisson processes with the same arrival rate, then the correlation attack will be ineffective. Consequently, it is sufficient and necessary to shape the traffic flows as Poisson processes with the same arrival rate in order to counter against the traffic analysis attacks based on these properties. In this case, the inter-packet intervals are exponentially distributed and packet arrival times are uniformly distributed. More discussion is left in Section IV.

Based on the above conclusion, what we need to do is to make every traffic flow have the same rate and shape the traffic flows to be Poisson processes with the same arrival rate, i.e., the same flow rate. As we assume an unobservable routing protocol is used, the flow rate that the adversary can observe is cumulated across concurrent unicast sessions. Meanwhile, it is desirable to make the transmission rate of as low as possible for energy efficiency. Thus, we formulate the following scheduling optimization problem:

$$\textbf{Privacy-Minimax:} \quad \min_{b,f} \; \max_{i} \; R_i \tag{6}$$

subject to:

$$\left\{ \begin{array}{rcll} \sum\limits_{j \in J} f_{iJj}^k & - & \sum\limits_{j \in \mathcal{N}} \sum\limits_{\{I | (j,I) \in \mathcal{A}, i \in I\}} f_{jIi}^k = \delta_i^k, \forall i, k, & \\ \sum\limits_k b_i^k & = & R_i & \\ \sum\limits_k b_i^k + \sum\limits_k \sum\limits_{j \in I(i)} b_j^k & \leqslant & C, & \forall i \neq S^k, \\ f_{iJj}^k & \leqslant & b_i^k p_{iJj}, & \forall i, \forall j \in J, \forall k. \end{array} \right. \tag{7}$$

where $i \in \mathcal{N}$, $k \in \mathcal{K}$ with $\mathcal{K}$ being the set of all concurrent sessions in the network, and

$$\delta_i^k = \left\{ \begin{array}{cl} r^k, & \text{if } i = S^k \\ -r^k, & \text{if } i = T^k \\ 0, & \text{otherwise.} \end{array} \right.$$

In this optimization problem, the objective is to minimize the maximum flow rate R_i for all nodes. Though it would not achieve the maximal privacy entropy as nodes' flow rate R_i's may be different, the difference between R_i and R_j for two nodes i and j is reduced. We can then inject padding traffic into each node to make R_i be the same. A more important result is this objective function tries to keep nodes' transmission rates low globally. The problem can be equivalently formulated as follows:

$$\textbf{Privacy-Minimax}^*\textbf{:} \quad \min_{b,f,t} \; R \tag{8}$$

subject to:

$$\left\{ \begin{array}{rcll} \sum\limits_{j \in J} f_{iJj}^k & - & \sum\limits_{j \in \mathcal{N}} \sum\limits_{\{I | (j,I) \in \mathcal{A}, i \in I\}} f_{jIi}^k = \delta_i^k, \forall i, k, & \\ \sum\limits_k b_i^k + \sum\limits_k \sum\limits_{j \in I(i)} b_j^k & \leqslant & C, & \forall i \neq S^k, \\ f_{iJj}^k & \leqslant & b_i^k p_{iJj}, & \forall i, \forall j \in J, \forall k, \\ \sum\limits_k b_i^k & \leqslant & R, & \forall i. \end{array} \right. \tag{9}$$

After a solution of the scheduling optimization problem is obtained, all nodes may have the same transmission rate in the ideal case. Then each node can transmit data in accordance to Poisson distribution, so that the adversary is not able to distinguish them. For the cases where not all nodes have the same transmission rate (i.e. $R_i = \sum\limits_k b_i^k$), we can simply inject dummy traffic at nodes to reach the maximum transmission rate R_i so they have the same transmission rate.

D. A Decentralized Network Scheduling Solution

Though the Privacy-Minimax* problem can be readily solved by standard linear programming algorithm, it is desirable to provide a decentralized solution for the network scheduling problem. In this section, we propose a decentralized algorithm for the network scheduling problem based on decomposition techniques [23]. Specifically, we decompose the original problem into three separate subproblems with decoupled variables based on the dual decomposition. Then we solve the subproblems independently, and finally solve the master dual problem by updating dual variables.

We first introduce Lagrange multipliers λ_i^k, μ_i, ν_{ij}^k, and φ_i to relax the four sets of constraints in (9) respectively. Then the Lagrangian function is as follows:

$$\begin{aligned}L(b,f,R,\lambda,\mu,\nu,\varphi) = R \quad &+ \quad \sum_i \sum_k \lambda_i^k (\sum_{j\in J} f_{iJj}^k \\ &- \quad \sum_{j\in\mathcal{N}} \sum_I f_{jIi}^k - \delta_i^k) \\ &+ \quad \sum_i \mu_i (\sum_k b_i^k + \sum_k \sum_{j\in I(i)} b_j^k - C) \\ &+ \quad \sum_i \sum_{j\in I(i)} \sum_k \nu_{ij}^k (f_{iJj}^k - b_i^k p_{iJj}) \\ &+ \quad \sum_i \varphi_i (\sum_k b_i^k - R)\end{aligned}$$

Note that in our model the interference node set $I(i)$ is equal to node i's end node set J. Thus, the original problem can be decomposed into three independent subproblems as follows:

$$\textbf{SUB1:}\ \ F_1 = \min_R (1 - \sum_i \varphi_i) R \tag{10}$$

$$\textbf{SUB2:}\ \ F_2 = \min_f \sum_i \sum_k \sum_{j\in J} f_{iJj}^k (\lambda_i^k - \lambda_j^k + \nu_{ij}^k) \tag{11}$$

$$\textbf{SUB3:}\ \ F_3 = \min_b \sum_i \sum_k b_i^k (\mu_i + \sum_{j\in I(i)} (\mu_j - \nu_{ij}^k p_{iJj}) + \varphi_i) \tag{12}$$

Since all the three subproblems are linear, the Lagrange multiplier method may not necessarily generate the optimal solution. We adopt the proximal method and add a quadratic term to make it strictly convex [24]. Take SUB2 as an example, the optimization problem to be solved at each node i is:

$$\min_f \sum_k \sum_{j\in J} f_{iJj}^k (\lambda_i^k - \lambda_j^k + \nu_{ij}^k)$$

Then we can add a quadratic term into it which turns the optimization problem to be:

$$\min_f \sum_k \sum_{j\in J} f_{iJj}^k (\lambda_i^k - \lambda_j^k + \nu_{ij}^k) + \frac{1}{2c} || f_{iJj}^k - f_{iJj}^k(t) ||^2$$

Then f_{iJj}^k is updated by

$$f_{iJj}^k(t+1) = [f_{iJj}^k(t) - c(\lambda_i^k - \lambda_j^k + \nu_{ij}^k)]^+$$

where c is positive constant scalar that makes the above update to be arbitrarily close to the optimal value of f_{iJj}^k, and $[\cdot]^+$ denotes the projection onto the non-negative orthant.

After the above three subproblems have been solved by each node for given $\lambda, \mu, \nu, \varphi$, we proceeds to solve the following master dual problem.

$$\max_{\lambda,\mu,\nu,\varphi} F_1 + F_2 + F_3 + \sum_i \sum_k \lambda_i^k \delta_i^k + \sum_i \mu_i C \tag{13}$$

subject to:

$$\mu_i > 0, \nu_{ij}^k > 0, \varphi_i > 0 \tag{14}$$

where F_1, F_2, F_3 are solutions to the subproblems for given $\lambda, \mu, \nu, \varphi$. The subgradient method can be used here to find the optimal solution to the master dual problem. In each iteration of subgradient optimization procedure, each Lagrange multiplier is updated according to its subgradient. For instance, λ_i^k is updated in each iteration by:

$$\lambda_i^k(t+1) = [\lambda_i^k(t) + \alpha(t)(\sum_{j\in J} \tilde{f}_{iJj}^k - \sum_{j\in\mathcal{N}} \sum_I \tilde{f}_{jIi}^k - \delta_i^k)]^+ \tag{15}$$

where t is the index of the iteration, $\tilde{f}_{iJj}^k$ is the optimal solution from subproblem SUB2, and $[\cdot]^+$ denotes the projection onto the feasible set of λ. $\alpha(t)$ is the step size for iteration t. A diminishing step size is adopted for the purpose of convergence. Specifically, $\alpha(t) = \frac{A}{1+B\cdot t}$ where A and B are non-negative tunable system parameters.

To summarize all the above procedures, we formulate the following decentralized Algorithm 1:

Algorithm 1 The Decentralized Privacy-Oriented Scheduling Optimization Algorithm

```
Input: Hypergraph: $(\mathcal{N}, \mathcal{A})$, Session Set: $\mathcal{K}$, Flow rate set: $\mathcal{R}$, Link reception probability set: $\mathcal{P}$, Link Capacity: $C$.
Output: $f_{iJj}^k$ and $b_i^k$.
  1) Initialization: set $t = 0$, $\lambda_i^k$ equal to some initial value, and $\mu_i, \nu_{ij}^k, \varphi_i$ equal to some non-negative values for all $i \in \mathcal{A}, j \in J, k \in \mathcal{K}$, where $(i, J) \in \mathcal{N}$.
  2) Each node $i$ locally solves its subproblems from SUB1, SUB2 and SUB3 for each session $k \in \mathcal{K}$, and then broadcast the result $f_{iJj}^k$ and $b_i^k$ to its direct neighbors, i.e., all nodes $j$ where $j \in J$ for $(i, J) \in \mathcal{A}$.
  3) Each node $i$ updates the Lagrange multipliers by the subgradient method as illustrated in (15). Then it broadcast $\varphi_i(t+1)$ to all other nodes, and broadcast $\lambda_i^k(t+1), \mu_i(t+1)$ to its direct neighbors.
  4) Set $t \leftarrow t+1$ and go to step 2 (until satisfying termination criterion).
```

It is important to note that it is unnecessary to broadcast $\nu_{ij}^k(t+1)$ to other nodes, and $\lambda_i^k(t+1), \mu_i(t+1)$ need to be broadcast to i's neighbors only. The convergence of the above algorithm follows the general convergence properties of subgradient and dual decomposition method.

After the optimal solution for the objective function is obtained, each node stores, encodes, and forwards packets towards the next hop with calculated transmission rate b_i^k at each session k. As the resulting network scheduling satisfies the flow constraints and capacity constraints, we conclude that the scheduling can achieve expected data rates.

In order to make each node have identical traffic pattern, we require each node to add appropriate dummy traffic. An obvious solution is to inject dummy traffic to the maximum transmission rate R_i. A good way to create dummy traffic is create redundant or linearly dependent packets, as these packets can improve reliability in lossy wireless transmission.

Then each node transmits packets in accordance to Poisson distribution.

IV. Traffic Analysis and Privacy Metric

In this section, we first demonstrate that traffic analysis attacks exploiting traffic patterns on inter-packet intervals, packet arrival time or data rates are ineffective in distinguishing data flows conforming to Poisson distribution with identical arrival rate. We do not consider content correlation attacks as in [7] since we assume that no content correlation information is leaked. Based on this result, we present an information theoretic metric on privacy for wireless networks. Then we use it to measure privacy quality offered by our scheme Priv-Code.

As assumed, there is an anonymous routing protocol which can provide unobservability [19] implemented in the wireless network. The adversary is not able to know the content of any packet, including the address field in the packet header.

A. Traffic Analysis

A node with an incoming flow conforming to Poisson distribution and exponential delay times can be viewed as an M/M/1 queuing system. According to Burke's theorem, the departure process of an M/M/1 queue is also a Poisson process with the same rate independent of the arrival process. This feature helps the proposed scheme Priv-Code thwart traffic analysis attacks. If incoming flows and outgoing flows are independent Poisson processes with the same arrival rate, it is impossible for an attacker to distinguish them.

Traffic analysis attacks have been proposed to exploit different traffic patterns of a flow, e.g., packet delay characteristic [20], number of packets in a fixed interval or window [21], [25].

Danezis [20] uses maximum likelihood estimation to distinguish different flows based on packet delay characteristics. Packets are delayed for a period conforming to the exponential distribution at the mixes, which is the optimal mixing strategy for a continuous-time mix network. Let X and Y are two output links of an exponential mix that the attacker wants to differentiate, C_X and C_Y are two model probability distributions for the two output links. Then the likelihood ratio can be formulated as

$$\mathcal{L} = \frac{\prod_{i=1}^{n} C_X(X_i) \prod_{j=1}^{m} u}{\prod_{i=1}^{n} u \prod_{j=1}^{m} C_Y(Y_j)},$$

where u is the uniform distribution parameter, and X_i and Y_j are sampled times coming out of channel X and Y respectively. It can be verified that if C_X and C_Y are uniform distributions then the likelihood ratio is 1, which is the case when the incoming traffic to the mix is a Poisson process. So we can see the attack fails when all traffic flows follow Poisson distribution with the same arrival rate.

Zhu et al.'s approach [21] correlates flows using number of packets in in a fixed interval. The pattern vector X_i of an input link or an output link i in Zhu et al.'s approach contains the following element:

$$X_{i,k} = \frac{\text{Number of packets in batch } k}{\text{Time elapses in batch } k}.$$

Then the mutual information between X_i and another pattern vector Y_j is

$$I(X,Y) = \int\int p(x_i, y_i) \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}.$$

This attack is especially effective against TCP due to the TCP loop-control mechanism. However, if the input flow and output flow conform to independent Poisson distribution as in our scheme, the attack will fail since the mutual information is actually 0.

A timing analysis attack proposed in [25] adopts a similar strategy as [21]. Specifically, for each possible entry-exit pair, the attacker computes the cross-correlation of the two sequences as

$$r(d) = \frac{\Sigma_i((x_i - \mu)(x_i' - \mu'))}{\sqrt{\Sigma_i(x_i - \mu)^2}\sqrt{\Sigma_i(x_i' - \mu')^2}},$$

where x_i is the number of packets received or sent by a mix during the ith window, μ and μ' are means of the two sequences. Correlation on the inter-packet intervals between two network links may lead to conclusion that they are carrying the same traffic. However, the cross-correlation is 0 for two independent Poisson processes, so the attack will not succeed in Priv-Code.

The adversary can launch two types of active attacks: artificial gaps and artificial bursts [25]. In the artificial gap attack, the adversary gets control of some valid nodes in the wireless network, and selectively drops several consecutive packets in a target flow to create a gap, which will result in a gap in other links. By examining the change on the inter-packet interval pattern, the adversary can identify related links. For artificial gaps, the attack must be prevented by injecting dummy traffic into nodes being influenced, so as to make the traffic pattern unchanged.

The artificial burst attack is to create a traffic burst by holding up packets at some nodes and release them at once. But such attacks can be effectively thwarted at each node by modulating outgoing data as a Poisson process for a predetermined rate. Moreover, the burst packets may be relayed through multiple paths, thus the burst is relieved by at each of the multiple links. As a result, the risk of one output link being found to be related to an input link is removed.

B. Information Theoretic Metric for Privacy

We adopt and extend the information theoretic approach for privacy measurement in [10], which proposed a privacy metric for mix networks. This metric is designed for mix networks, which have several differences with the network model used in our scheme. The main difference is that a node in our network model can be a sender or a receiver when it works as a "mixing" relay for others, while a mix in a typical mix network is normally not a sender or a receiver. Another difference is that nodes in in our network model may form a loop, which does not exist in the mix network. Hence, we have to adapt the information theoretic metric for privacy to our network model.

The privacy metric defined in [10] uses entropy to describe privacy quality. Specifically, sender anonymity (or receiver anonymity) is the entropy of the attacker's probability distribution of users being the sender (or receiver) with respect to a message. Let Ψ be the set of all users, $u \in \Psi$ be the user, and p_u be the probability of the user u being the sender of a message M. Then the privacy measurement, called the effective anonymity size, of the sender anonymity with respect to M is:

$$H(M) = -\sum_{u \in \Psi} p_u \cdot \log p_u.$$

The privacy metric can be interpreted as the number of bits of additional information that the attacker needs to identify the user u being the message sender. It is trivially to see that if $p_u = 1$ for some user u then the entropy is 0 bits, meaning the attacker has identified the user already.

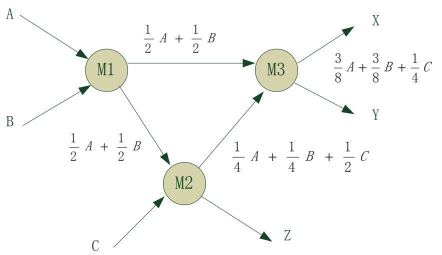

A
B
C
M1
M2
M3
X
Y
Z
$\frac{1}{2}A + \frac{1}{2}B$
$\frac{1}{2}A + \frac{1}{2}B$
$\frac{1}{4}A + \frac{1}{4}B + \frac{1}{2}C$
$\frac{3}{8}A + \frac{3}{8}B + \frac{1}{4}C$

Fig. 1. A Simple Mix Network and Its Privacy Metric. The sender anonymity of the output links of mix 1 is $H_1^{total} = -(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) = 1$, and the sender anonymity of the output links of mix 3 is $H_3^{total} = -(\frac{3}{8}\log\frac{3}{8} + \frac{3}{8}\log\frac{3}{8} + \frac{2}{8}\log\frac{2}{8}) = 1.56$.

Figure 1 illustrates a simple mix network consisting of 3 mixes, 3 senders and 3 receivers. Assume the attacker does not have any a-priori knowledge about the senders and receivers except the traffic pattern. When all message flows are Poisson processes with the same rate, a message flow having arrived at a mix was equally likely to have been forwarded to all of the possible "next hops". Then the probability distribution of the output link of a mix forwarding input data flows is showed in the figure. For example, the probability distribution of output links of mix 2 is $\{A : \frac{1}{4}, B : \frac{1}{4}, C : \frac{1}{2}\}$. We rewrite the probability distribution as $\{\frac{1}{4}A + \frac{1}{4}B + \frac{1}{2}C\}$. Thus we can compute the privacy entropy coming out of mix node 3 is:

$$H_3^{total} = -(\frac{3}{8}\log\frac{3}{8} + \frac{3}{8}\log\frac{3}{8} + \frac{2}{8}\log\frac{2}{8}) = 1.56,$$

which means the attacker needs 1.56 bits information to identify who is the sender. Note that both mix 1 and mix 2 have contributed to the privacy entropy besides mix 3, and the final privacy entropy is the synergistic effect of all three mixes.

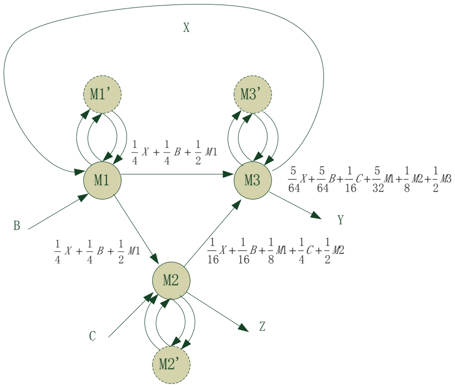

X
M1'
M3'
M1
M3
M2
M2'
B
C
Y
Z
$\frac{1}{4}X + \frac{1}{4}B + \frac{1}{2}M1$
$\frac{5}{64}X + \frac{5}{64}B + \frac{1}{16}C + \frac{5}{32}M1 + \frac{1}{8}M2 + \frac{1}{2}M3$
$\frac{1}{4}X + \frac{1}{4}B + \frac{1}{2}M1$
$\frac{1}{16}X + \frac{1}{16}B + \frac{1}{8}M1 + \frac{1}{4}C + \frac{1}{2}M2$

Fig. 2. The Extended Network for Privacy Measurement. For each node M, an auxiliary node M' is created so as to support the cases where the mix nodes can be senders or receivers. It also contains a loop from M1 to M3.

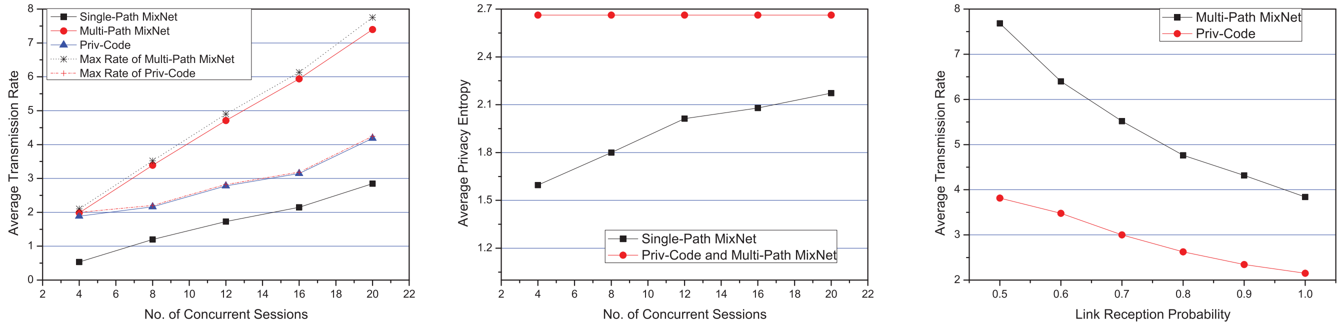
We extend the privacy metric to the network model where the mixes can be senders or receivers and they can form message flow loops. Fig. 2 extends the simple mix network in Fig. 1 in two aspects. First, we introduce an auxiliary node into the graph for each mix, and adds the same number of incoming and outgoing edges between the auxiliary node and the mix node. Then the resulting graph is able to correctly describe cases in which the mix nodes are senders or receivers. For instance, the new graph is able to describe two sessions between B to M1 and M1 to M3 by $B \to M1 \to M1'$ and $M1' \to M1 \to M3 \to M3'$. Next, if there is a message flow loop from M3 to M1 in the graph as showed in Fig. 2, then we need to determine the probability of the message flow over the loop. In this case, we assume the message flow from M3 to M1 is X, then we can infer the message flow over every link as shown in Fig. 2. Since the message flow from M3 to M1 is X, we can have $\frac{5}{64}X + \frac{5}{64}B + \frac{1}{16}C + \frac{5}{32}M1 + \frac{1}{8}M2 + \frac{1}{2}M3 = X$, yielding $X = \frac{1}{59}(5B + 4C + 10M1 + 8M2 + 32M3)$. Then the privacy entropy of each flow can be obtained by the formula above.

Then we can use this privacy measurement approach to evaluate the proposed scheme Priv-Code. Suppose the wireless network consists of N nodes and each node is transmitting data as an independent Poisson process of the same transmission rate. Then the privacy entropy of an output data flow of a node can be computed from the probability distribution of the input flows.

V. Performance Analysis and Simulation

Computation Overhead: The computation overhead comes from two sources, the encoding procedure and the anonymous routing protocol. Computation cost of the anonymous routing protocol is relatively lightweight, since there is very few public key operations in the protocol. Hence we mainly focus on the encoding computation overhead.

The proposed scheme does not rely on expensive public key cryptographic mechanisms to protect encoding vectors, either. At each source/intermediate node, a random coding



(a) Average Transmission Rate of Nodes vs. Session Number

(b) Average Privacy Entropy vs. Session Number

(c) Average Privacy Entropy vs. Link Reception Probability

Fig. 3. Privacy Entropy and Transmission Rate Comparison of Priv-Code, Single-Path and Multi-Path Mix Network. The Network size is 50, each node has 5 neighbors on average, link capacity is 100 units, and each communication session is 2 units.

vector is generated and used to encode packets from the same session. Then the coding vector of the newly generated packet is attached to the packet header. This procedure is much more efficient than encoding/decoding using homomorphic encryption.

For each packet received at an intermediate node, it needs to verify whether it is independent from cached packets. *Gauss-Jordan elimination* can be used to check whether a packet is linearly independent, and the computation complexity is $O(n^3)$, where n is the size of a generation.

Communication Overhead: For network coding, the encoding coefficients need to be put in the packet header for the intermediate nodes or the destination node to re-encode or decode packets. This part of overhead results in additional communication cost. If the coefficients obtain their value from 0 to 255, i.e., the size of a byte, then the size of coefficients in a packet header is n bytes. If the generation size is 20, that is, 20 packets are grouped into a generation, then the network coding coefficient overhead is 20 bytes. For a packet with size of 1000 bytes, the overhead is only 2% of the whole packet size.

Storage Requirement: In the proposed scheme, the source node has to cache all packets in the generation before an acknowledgement is received; the intermediate nodes need to cache all received linearly independent packets in the generation before receiving an acknowledgement; the destination also has to cache all linearly independent packets in the same generation before they can be decoded. Thus, the network coding mechanism demands much more storage space than traditional transmission technique. If the size of a generation is n, and on average there are m concurrent sessions passing through a node, then a node has to allocate $O(nm)$ packet cache on average. Note that the intermediate nodes may not need nm packet cache since the packets in a generation are transmitted via multiple paths. This requirement on storage can be tuned by setting the generation size n, and is also determined by the network traffic.

Simulation and Evaluation: We implement our proposed network coding scheduling scheme and conduct experiments with MatLab and NS2 to evaluate its privacy protection capability and performance. We use the privacy metric presented in Section IV to compare our scheme Priv-Code with two typical mix networks, a single-path mix network and a multi-path mix network. The single-path mix network always chooses the shortest path from a source to its destination, while the multi-path mix network selects multiple paths to transmit data and it uses a similar optimization approach as Priv-Code. The multi-path mix network is different from Priv-Code in that transmissions over paths are independent of each other. Both mix networks for comparison are composed of exponential mixes that delay packets according to the exponential distribution, so the output flows are independent Poisson processes.

The wireless network in our experiments consists 50 randomly deployed nodes with node density 6, i.e., each node has 5 neighbors on average [7]. We assume that all links have the same capacity, and nodes within the interference range share the capacity. The link capacity must guarantee the optimization algorithm has a optimal solution. In our experiments, we fix the link capacity to be 100 units.

In our experiments, we change the following parameters in our experiments:

- Concurrent session number: The number of concurrent sessions ranges from 4 to 20, which means there are at most 20 pairs of nodes are communicating at the same time. The communication rate of each session is 2 units.
- Link reception probability: The reception probability of each link p_{iJj} is a tunable parameter, whose distribution conforms to the uniform distribution with a mean ranging from 0.5 to 1.0.

We compute the average privacy entropy of all message flows based on the privacy measurement method in Section IV, and at the same time, we compute the average transmission rate and the maximum transmission rate among all nodes in the network. Fig. 3(a) shows the average transmission rate of the single-path mix network, the multi-path mix network and our scheme Priv-Code. Since we have to inject dummy traffic into the network to make each node have identical transmission rate, i.e. the maximum transmission rate, we also show the maximum transmission rate for each experiment in this figure. It can be seen that the required transmission rate of Priv-Code is between the single-path mix network and the multi-path mix network, which accords with our expectation as Priv-Code tries to provide both strong privacy and good performance. It means that Priv-Code can transmit the given traffic with less transmission rate than the multi-path mix network. Note that dummy traffic of Priv-Code to be injected into the network is also less than that of the multi-path mix network.

We show in Fig. 3(b) the average privacy entropy over all nodes provided by Priv-Code, the single-path mix network and the multi-path mix network. Since we make each node have identical transmission rate by traffic padding, Priv-Code and the multi-path mix network have the same privacy entropy. In contrast, the single-path mix network always chooses the shortest path for data transmission, the privacy changes with the number of concurrent sessions in the network. When there are fewer sessions, it is harder for the single-path mix network to protect their privacy, but the privacy entropy grows as the number of sessions increases.

Fig. 3(c) shows the average transmission rate versus different link reception probability for Priv-Code and the multi-path mix network. It shows that the required transmission rate of both Priv-Code and the multi-path mix network decreases as the link quality becomes better. But the transmission rate of Priv-Code is almost half of that of the multi-path mix network, which means Priv-Code can transmit the same amount of traffic with about half of the transmission rate compared with the multi-path mix network. demonstrates the great advantage of network coding in performance improvement.

VI. Conclusion

In this paper, we investigate the problem of how to exploit network coding to protect privacy against traffic analysis attacks under a powerful threat model, in which the attacker is able to continuously monitor the entire network and mount both passive and active attacks against the wireless network. Based on information theory, we formally define the privacy entropy in terms of traffic flow information. Then we formalize a hypergraph-based network model for wireless networks based on network coding, and formulate an optimization problem to seek the optimal network coding scheduling. After that, we provide a decentralized algorithm to solve the optimization problem. Our analysis and experiment evaluation show that the proposed scheme has substantial advantage over existing schemes on privacy protection against traffic analysis.

Acknowledgments

Zhiguo's study is supported in part by Scientific Foundation for Returned Overseas Chinese Scholars, MOE, and the NSFC project under Grant No. 61003223. Kai's research is supported by NSFC under grant 61170267, Jiangsu Natural Science Foundation under grant BK2011358, and National Basic Research Program of China (973) under grant 2011CB302905. Yunhao's study is supported in part by NSFC Distinguished Young Scholars Program under Grant 61125202, and National High-Tech R&D Program of China (863) under grant No. 2011AA010100.

References

[1] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 4, no. 2, February 1981.

[2] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for web transactions," *ACM Trans. Inf. Syst. Secur.*, vol. 1, no. 1, pp. 66–92, 1998.

[3] R. Dingledine, N. Mathewson, and P. F. Syverson, "Tor: The second-generation onion router," in *USENIX Security Symposium*, 2004, pp. 303–320.

[4] U. Moller and L. Cottrell, "Mixmaster protocol–version 2," IETF Internet draft, 2003.

[5] G. Danezis, R. Dingledine, and N. Mathewson, "Mixminion: Design of a type iii anonymous remailer protocol," in *IEEE Symposium on Security and Privacy*, 2003, pp. 2–15.

[6] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[7] Y. Fan, Y. Jiang, H. Zhu, and X. Shen, "An efficient privacy-preserving scheme against traffic analysis attacks in network coding," in *INFOCOM*, 2009, pp. 2213–2221.

[8] T. Cui, L. Chen, and T. Ho, "Energy efficient opportunistic network coding for wireless networks," in *INFOCOM*, 2008, pp. 361–365.

[9] G. Danezis and I. Goldberg, "Sphinx: A compact and provably secure mix format," in *IEEE Symposium on Security and Privacy*, 2009, pp. 269–282.

[10] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Privacy Enhancing Technologies*, 2002, pp. 41–53.

[11] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, and J. Crowcroft, "Xors in the air: practical wireless network coding," in *SIGCOMM*, 2006, pp. 243–254.

[12] M. N. Krohn, M. J. Freedman, and D. Mazières, "On-the-fly verification of rateless erasure codes for efficient content distribution," in *IEEE Symposium on Security and Privacy*, 2004, pp. 226–240.

[13] Z. Yu, Y. Wei, B. Ramkumar, and Y. Guan, "An efficient scheme for securing xor network coding against pollution attacks," in *INFOCOM*, 2009, pp. 406–414.

[14] X. Chang, J. Wang, J. Wang, V. Lee, K. Lu, and Y. Yang, "On achieving maximum secure throughput using network coding against wiretap attack," in *ICDCS*, 2010, pp. 526–535.

[15] J. Wang, J. Wang, K. Lu, B. Xiao, and N. Gu, "Optimal linear network coding design for secure unicast with multiple streams," in *INFOCOM*, 2010, pp. 2240–2248.

[16] S. Jaggi, M. Langberg, S. Katti, T. Ho, D. Katabi, M. Médard, and M. Effros, "Resilient network coding in the presence of byzantine adversaries," *IEEE Transactions on Information Theory*, vol. 54, no. 6, pp. 2596–2603, 2008.

[17] P. Zhang, Y. Jiang, C. Lin, Y. Fan, and X. Shen, "P-coding: Secure network coding against eavesdropping attacks," in *INFOCOM*, 2010, pp. 2249–2257.

[18] Y. Jiang, Y. Fan, X. S. Shen, and C. Lin, "A self-adaptive probabilistic packet filtering scheme against entropy attacks in network coding," *Computer Networks*, vol. 53, no. 18, pp. 3089–3101, 2009.

[19] Z. Wan, K. Ren, B. Zhu, B. Preneel, and M. Gu, "Anonymous user communication for privacy protection in wireless metropolitan mesh networks," *IEEE Trans. Veh. Technol.*, no. 2, pp. 519 – 532, 2010.

[20] G. Danezis, "The traffic analysis of continuous-time mixes," in *PET04*, 2004.

[21] Y. Zhu, X. Fu, B. Graham, R. Bettati, and W. Zhao, "On flow correlation attacks and countermeasures in mix networks," in *PET04, LNCS 3424*, 2004, pp. 207–225.

[22] X. Zhang and B. Li, "Optimized multipath network coding in lossy wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 5, pp. 622–634, 2009.

[23] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[24] D. P. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

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