Enabling Network Anti-Inference via Proactive Strategies: A Fundamental Perspective

Zhuo Lu, *Member, IEEE,* and Cliff Wang, *Fellow, IEEE,*

***Abstract*—Network inference is an effective mechanism to infer end-to-end flow rates and has enabled a variety of applications (e.g., network surveillance and diagnosis). The paper is focused on the opposite side of network inference, i.e., how to make inference inaccurate, which we call *network anti-inference*. As most research efforts have been focused on developing efficient inference methods, design of anti-inference is largely overlooked. Anti-inference scenarios can rise when network inference is not desirable, such as in clandestine communication and military applications. Our objective is to *explore network dynamics to provide anti-inference*. In particular, we consider two proactive strategies that cause network dynamics: *transmitting deception traffic and changing routing to mislead the inference*. We build an analytical framework to quantify the induced inference errors of the proactive strategies that maintain limited costs. We find via analysis and simulations that for deception traffic, a simple random transmission strategy can achieve inference errors on the same order of the best coordinated transmission strategy; while changing routing can cause inference errors of higher order than any deception traffic strategy. Our results not only reveal the fundamental perspective on proactive strategies, but also offer the guidance into practical design of anti-inference.**

***Index Terms*—Proactive strategies, network inference and tomography, network dynamics, security.**

I. Introduction

Network inference, also known as network tomography, is an effective way to infer end-to-end flow or link rates from network measurements [1]–[8]. The essential idea of network inference is to formulate a relationship (determined by the routing protocol) between end-to-end flow and link rates, then infer via such a relationship. Network inference has enabled a wide range of applications, such as network surveillance, management and diagnosis [1], [2], [4], [8], [9].

In this paper, we focus on the opposite side of network inference, i.e., how to make inference inaccurate, which we name as *network anti-inference*. We aim to advance the anti-inference strategies, which have not yet been fully studied. Our work is motivated by scenarios where network inference is not desirable or even malicious. For example, in military applications, nodes may want to hide end-to-end flow rates from the adversaries that attempt to infer such information. If the adversaries obtain the information, they can know who is communicating with whom in the network; and more importantly, they can further infer who is a commanding node that sends the same/similar commanding data (with the same/similar rate) to a few other nodes under the Command and Control (C2) orders [10]. Thus, an anti-inference strategy can be a vital solution to make sure the flow rates in the network are never correctly inferred.

As network inference is based on inferring via the relationship between flow and link rates, there are two immediate strategies to offer anti-inference: (i) transmitting redundant traffic called *deception traffic* into the network to cause substantial inference errors, and (ii) keeping changing routing such that the attacker cannot correctly acquire the relationship that varies over time. Both strategies are proactive; i.e., they must be deployed and executed to prevent inference, and can potentially degrade the network performance.

As security usually comes with a cost, the key question for a security measure is how much benefit can be obtained under a reasonably limited cost. For example, it is highly desirable if a commander's communication flows to soldiers cannot be accurately identified by transmitting an inconsiderable amount of deception traffic to mislead the attacker. Therefore, our objective in this paper is to *understand the fundamental impact of proactive strategies with a bounded cost for network anti-inference*. In particular, we consider a wireless network in the presence of an attacker that attempts to infer all end-to-end flow rates via eavesdropping on network links for its malicious purpose. We focus on investigating the impact of two proactive methods: deception traffic and routing changing. We build an analytical framework to quantify what impacts (in terms of inference errors) the two strategies can bring while maintaining a limited cost, such as slight throughput or delay degradation. We use simulations to evaluate the impacts of proactive strategies in practical network inference setups. To the best of our knowledge, we are the first to systematically study the proactive strategies for network anti-inference. The major findings and contributions are summarized as follows.

- We found that for the deception traffic strategy that causes a limited performance degradation, independently transmitting random traffic at each node can cause the inference error on the same order of the best coordinated transmission strategy in all nodes. Further, the inference error will be increased by at least a few order of magnitude if the mean rate of the deception traffic is kept a secret from attackers.
- We discovered that under a constant delay degradation, proactively changing routing paths in general leads to the inference error of higher order of magnitude than any deception traffic strategy.

Zhuo Lu is with the Department of Computer Science, University of Memphis, TN. E-mail: zhuo.lu@memphis.edu.

Cliff Wang is with Army Research Office, Research Triangle Park, NC. Email: cliff.wang@us.army.mil.

An earlier version of this work was presented in IEEE Conference on Computer Communications (INFOCOM) 2015. Zhuo Lu is supported by NSF CNS-1464114.

- We showed that combining deception traffic and routing changing cannot significantly boost the impact of anti-inference. Rather, the induced inference error is dominated by whatever individual strategy that leads to more error than the other. This means that the combined strategy is not always desirable because of its slight improvement of anti-inference at the double cost (i.e., redundant traffic and potentially non-optimal routing change).

Our results reveal the fundamental perspective of exploring network dynamics to provide defense against network inference. The findings in this paper can not only show the impact region of a proactive strategy for a network scenario, but also provide the performance benchmark and guidance for design of anti-inference protocols for practical use.

The rest of this paper is organized as follows. In Section II, we introduce models and network inference. In Sections III and IV, we present our findings and prove the results, respectively. Then, in Section V, we discuss the simulation results. Section VI, we present related work. Finally, we summarize the conclusions in Section VII.

II. Preliminaries and Problem Statement

In this section, we introduce models and assumptions, then state the research problem of network anti-inference. Notations: We write $f(x) = O(g(x))$ or $g(n) = \Omega(f(n))$ if $\exists\ n_0 > 0$ and constant c_0 such that $f(n) \leq c_0 g(n)\ \forall n \geq n_0$. We write $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(n) = \Omega(g(n))$. We denote by $\mathbf{A}^T$ the transpose of matrix $\mathbf{A}$. The $\mathcal{L}_1$ norm of vector $\mathbf{a} = [a_1, a_2, \cdots, a_k]^T$ is defined as $\|\mathbf{a}\|_1 = \sum_{i=1}^k |a_i|$. Similarly, the $\mathcal{L}_2$ norm is $\|\mathbf{a}\|_2 = \sum_{i=1}^k a_i^2$. We denote by $\mathbf{tr}\{\mathbf{A}\}$ the trace of matrix $\mathbf{A}$, i.e., the sum of all its eigenvalues.

A. Network Models

There are many network models used for the network inference problem in the literature, such as random power-law graphs, Erdos-Renyi graphs, and random geometric graphs [11]. We choose the random geometric graph as our model, as it has been widely-adopted for wireless network study. We consider a wireless network with n nodes distributed independently and uniformly on region $\Omega = [0, \sqrt{n/\lambda}]^2$ for a large node density λ such that the network is connected (asymptotically almost surely) [12]. We say two nodes have a network link if they are in each other's transmission range r.

In the network with n nodes, there are at most $n(n-1)/2$ end-to-end flows if links are undirected, or $n(n-1)$ flows if links are directed. We assume that all links are undirected, since the directed case is a straightforward extension to the undirected one and the assumption does not affect the formulation of the inference/anti-inference problem. We denote by L the total number of undirected links in the network.

We assume that each node has at most a finite number of end-to-end flows to other nodes in the network. In other words, there are $F = O(n)$ end-to-end flows in the network.

B. Attack Model and Network Inference

There exists an attacker attempting to use network inference to infer all end-to-end flow rates. We make following assumptions on the attacker.

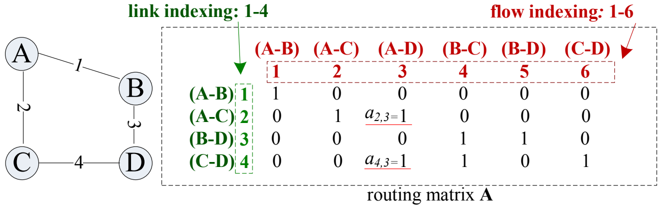
- We assume that the attacker has the capability of overhearing all link transmissions (e.g., by placing eavesdroppers all over the network). However, the attacker can only use physical/link layer information, and cannot access transport layer or higher layer information since it is usually encrypted in data packets (e.g., [13], [14]).
- The attacker is aware of the network topology; hence, given a routing protocol used in the network (e.g., shortest path routing), the attacker knows the routing path for any flow. We do not consider node mobility and link stability; thus the topology does not change over time.

Given the attacker's capability, we describe how it infers all flow rates. First, there are at most $n(n-1)/2$ potential flows in the network, which can be indexed from 1 to $n(n-1)/2$. Apparently, the attacker does not exactly know which flows indeed exist. However, on the other hand, it is equivalent to say that the rate of a flow is zero if the flow does not exist. Therefore, the attacker's inference strategy is to assume that there exist $n(n-1)/2$ flows, in which some of them have zero rate. In this way, all $n(n-1)/2$ flows are associated with a flow rate vector $\mathbf{x} \in \mathbb{R}^{(n(n-1)/2)\times 1}$, whose entry represents the rate of each flow. The goal of the attacker is to obtain an estimate $\hat{\mathbf{x}}$ in close value to $\mathbf{x}$. However, the attacker cannot directly see $\mathbf{x}$, but can only observe the data transmission on each link. This means that the attacker can obtain the observed link rate vector as $\mathbf{y} \in \mathbb{R}^{L\times 1}$ (as there are L links in the network), whose entry is the data transmission rate at each link.

It has been shown [1], [3], [5]–[7] that $\mathbf{x}$ and $\mathbf{y}$ have a linear relationship, i.e.,

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{A} = \{a_{i,j}\}$ is an $L \times \frac{n(n-1)}{2}$ matrix, called the routing matrix in the network, whose element $a_{i,j}$ has value 1 if the i-th link is on the routing path of flow j, and value 0 otherwise.



		(A-B)	(A-C)	(A-D)	(B-C)	(B-D)	(C-D)
		1	2	3	4	5	6
(A-B)	1	1	0	0	0	0	0
(A-C)	2	0	1	$a_{2,3}=1$	0	0	0
(B-D)	3	0	0	0	1	1	0
(C-D)	4	0	0	$a_{4,3}=1$	1	0	1

Fig. 1. Example in a four-node network: how to build up the routing matrix.

In the following, we use a toy example to show how the routing matrix $\mathbf{A}$ is determined. In Fig. 1, there are only 4 nodes A, B, C, D and 4 undirected links 1 (A-B), 2 (A-C), 3 (B-D), 4 (C-D) in the network. There can be 6 potential end-to-end flows in the network: 1 (A-B), 2 (A-C), 3 (A-D), 4 (B-C), 5 (B-D), and 6 (C-D). Note that all links and potential end-to-end flows are indexed (starting from 1) in network inference.

The routing matrix $\mathbf{A}$ is a 4-by-6 matrix representing how point-to-point links form end-to-end flows. In particular, $a_{i,j}$ is 1 if the j-th flow is routed over the i-th link, and is 0 otherwise.

For example in Fig. 1, flow 3 (A-D) will be routed over link 2 (A-C) and link 4 (C-D). Therefore, we can see that $a_{1,3} = 0$, $a_{2,3} = 1$, $a_{3,3} = 0$, and $a_{4,3} = 1$ in $\mathbf{A}$. Now suppose that only node A has a data flow with 100bps to D. Then, the attacker can observe on links 2 and 4 that there are data transmissions with rate 100bps. Therefore, the goal of the attacker is to infer flow rate vector (with true value $\mathbf{x} = [0, 0, 100, 0, 0, 0, 0]^T$) from link observation vector $\mathbf{y} = [0, 100, 0, 100]^T$.

It is obvious that things become complicated if there are more nodes and flows. How can the attacker infer all end-to-end flow rates from the observations on each link? It has been shown that (1) is usually an under-determined system (e.g., there are four links and six flows in Fig. 1), thus the conventional least squares estimation cannot be directly applied.

There is a line of work (e.g., [1]–[5], [7], [15], [16]) that has already studied this problem. It has been shown that if the end-to-end flow vector $\mathbf{x}$ is sparse (as it is less likely that everyone is communicating with everyone in practice), it can be recovered (with high probability) from the under-determined system $\mathbf{y} = \mathbf{A}\mathbf{x}$ using $\mathcal{L}_1$-norm minimization

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^{(n(n-1)/2) \times 1}}{\arg\min} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1, \qquad (2)$$

The $\mathcal{L}_1$-norm minimization (2) has been widely adopted and extended to infer network statistics in many applications (e.g., [6], [7], [17], [18]).

Note that the $\mathcal{L}_1$-norm minimization and its variants can be viable solutions for network flow inference. An attacker may attempt to use any techniques, including but not limited to $\mathcal{L}_1$-norm minimization, to infer network flow information. If we know what technique the attacker uses, we may design specific anti-inference strategy against such a technique. To make our results general, we assume that we do not know what technique the attacker may use for network inference. This general assumption will not limit our anti-inference results in the scope of a specific inference method.

The goal of the attacker in this paper is to infer some or all end-to-end flow rates. However, even when it wants to infer some but not all rates, it still has to formulate the inference problem globally to get the rates of these desirable flows via first formulating all flow rates. This is because end-to-end flows can cross paths with each other and thus are coupled in the inference formulation in (1). To obtain some rates in the vector $\mathbf{x}$ in (1), we still need to solve the general inverse problem to obtain the estimate of $\mathbf{x}$ in (1) first.

It is also worthy of mentioning that this paper considers a strong attack model in which the attacker can monitor all links, thereby corresponding to the worst case. This is because it is desirable to assume a powerful attack model when analyzing the effectiveness of countermeasure design. Thus, we adopt the worst-case analysis in this paper. In fact, even if the attacker can only monitor some (but not all) links in the network, our formulation (1) can be easily adapted by removing the link entries in $\mathbf{y}$ that the attacker cannot observe. The follow-up analysis can be also adapted in a similar way.

C. Anti-Inference Problem

As aforementioned, anti-inference is to make inference inaccurate. To this end, we take a close look at the relationship between flow rates and observations in (1), and find two major factors that can affect inference.

1) The observation vector $\mathbf{y}$ depends on what nodes transmit. It is evident that a node must transmit the data that it should do. Thus, in order to make an impact on inference, the node can transmit redundant traffic for the deception purpose, causing observation errors in $\mathbf{y}$. We refer to such traffic as deception traffic.
2) The routing matrix $\mathbf{A}$ is determined by a routing protocol. If a node deliberately selects a routing path that is not predictable to the attacker, it will cause a routing matrix mismatch in (1) and lead to inference error.

This means that we can either transmit deception traffic or change routing to offer anti-inference. However, both methods come with penalty: transmitting deception traffic makes the network more congested; and changing routing can degrade the performance (e.g., end-to-end delay). As enhancing security usually brings costs, a fundamental and key question is how much benefit we can get if we limit the costs of such proactive strategies. In the next section, we aim to answer this question by quantifying the benefit of network anti-inference under limited costs.

It is worth mentioning that deception traffic strategies have been used for clandestine communication [13], [14], [19]. However, the main scope of these methods is to make traffic transmissions look like independent on a particular end-to-end path without considering the global network traffic pattern. The deception traffic strategy in this paper aims to lead to errors in inference based on the global network view.

III. Analyzing Benefits and Costs of Anti-Inference

As the attacker can choose any method (e.g., $\mathcal{L}_1$-norm based [7], [20]) to infer the network flows, the error induced by a proactive strategy hinges on the inference method that the attacker uses. If we know what specific method the attacker chooses to infer, we may design an anti-inference strategy accordingly against such an inference method. However, the inference attacker is always passive and is not involved in any network activity; therefore it is quite difficult to know what exact inference method it may use. To provide a fundamental view on proactive strategy based anti-inference, we aim at modeling the impact of proactive strategies on the genie bound of network inference, which represents the inference error achieved by the theoretically best inference method. Our goal is to see how much proactive strategies can increase such a genie bound (thereby causing more error for any inference method that the attacker may choose).

In this section, we first define the genie bound of network inference, then present and discuss the main results of benefits and costs of anti-inference. We will prove all main results in Section IV.

A. The Genie Bound

The genie bound is a lower bound [15], [16] of errors to solve an under-determined linear system like (1). It denotes the optimal performance among all possible methods and is derived in two steps: first, assume that there is a genie that tells us who is actually having an end-to-end flow to whom in the network; then, based on such information and given observations, the least squares estimate is derived to minimize the mean square error of flow rate estimation. In particular, the genie bound is derived by measuring the error between the true flow rate and the least squares solution to a new equation formed by directly removing zero entries in $\mathbf{x}$ (assisted by the genie) in (1). In this way, the attacker's use of a particular inference method, such as the $\mathcal{L}_1$ minimization in (2), will have no effect on computing the genie bound, which is a general, method-independent bound, serving as the lower error bound for any inference method.

In our context, the genie bound of network inference under a proactive strategy is defined as follows.

Definition 1: When the number of real flows F is no greater than the number of links L in the network, the genie bound conditioned on a proactive strategy $\mathcal{S}$ is the minimum mean square error of traffic rate estimation for all end-to-end flows in the network, i.e.,

$$\mathcal{G}(\mathbf{x}_g|\mathcal{S}) = \mathbb{E}\left(\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2|\mathcal{S}\right),$$

where $\mathbf{x}_g \in \mathbb{R}^{F\times 1}$ is the flow rate vector for node pairs that indeed have end-to-end flows and $\hat{\mathbf{x}}_g$ is the minimum mean square error estimate of $\mathbf{x}_g$.

Given Definition 1, we have a metric to quantify how proactive strategies can badly affect network inference. Note that a necessary condition to compute the genie bound is that $F \leq L$. Therefore, we assume that $F \leq L$ in the rest of the paper and aim at understanding how proactive strategies can increase the genie bound of network inference based on the formulation in (1). To proceed, we first need to discuss the statistical property of the routing matrix $\mathbf{A}$ in (1), which will be extensively used to later analysis.

B. Routing Matrix Modeling

As a key component in (1), the routing matrix $\mathbf{A}$ is a random matrix because nodes are randomly distributed over the network region. Moreover, matrix $\mathbf{A}$ depends on the routing protocol used in the network. Hence, it is non-trivial to characterize matrix $\mathbf{A}$ under any (class of) routing protocol(s), or under any routing changing strategy. In the following, we propose an important technical model for the routing changing strategy to serve as a mathematically tractable yet generic model to tackle the anti-inference problem.

Model 1: Under any routing strategy considered in this paper, the average number of hops between any source-destination pair is denoted by a function $g(n)$ satisfying $g(n) = O(n)$, where n is the number of nodes in the network.

Remark 1: Model 1 technically limits our scope into a set of certain routing protocols. In essence, it states that a reasonable routing protocol should on average give a path with a limited number (no higher order of n) of forwarding nodes. It can be verified that a wide range of practical routing models, such as the K-shortest path routing, belong to Model 1. For example in Fig. 2, under Model 1, we only consider routing protocols that find any path illustrated in solid lines, and exclude protocols that give a much longer path (e.g., the one in dotted line).

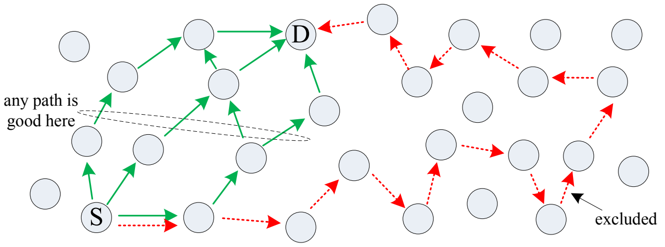


Fig. 2. Node selections in routing protocols from source S to destination D.

With Model 1, we are in fact able to model the statistical characteristics of the routing matrix $\mathbf{A}$. In particular, we have the following lemma.

Lemma 1: Under Model 1, the probability that element $a_{i,j}$ in routing matrix $\mathbf{A}$ is 1 is

$$\mathbb{P}(a_{i,j} = 1) = \Theta\left(\frac{g(n)}{n}\right). \tag{3}$$

Proof: See Appendix. □

Remark 2: Lemma 1 shows that statistically, a routing path includes a particular link with probability on the order of $g(n)/n$ in the network. This indicates that if a routing protocol has a longer routing path (i.e., when $g(n)$ becomes larger), a link will be more likely to be selected to forward data, which is also intuitively true. Lemma 1 provides statistical properties of different routing (or routing-changing) protocols that will enable us to analyze the impacts of anti-inference strategies.

C. Analyzing Deception Traffic

After defining the genie bound and the routing model, we can quantify the benefits and costs of anti-inference strategies. We first investigate how transmitting deception traffic helps prevent end-to-end flows from being inferred. We consider the proactive strategy that each node transmits deception traffic to its one-hop neighbor.

Let $\mathbf{J} = [J_1, J_2, \cdots, J_L]^T$ be the deception traffic rate vector, where J_i is the deception traffic rate for the i-th link ($i \in [1, L]$) in the network. Then, compared with the original formulation in (1), the link rate vector $\mathbf{y}$ (observed by the attacker) due to deception traffic can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{J}, \tag{4}$$

where the deception traffic rate vector $\mathbf{J}$ can be considered as a noise vector in the linear formulation. We want to know what the genie bound is for the attacker who wants to infer the flow rate vector $\mathbf{x}$ from the observed link rate vector $\mathbf{y}$.

In the following, we consider two cases: 1) nodes transmit deception traffic in a coordinated way, and 2) nodes independently transmit random deception traffic.

1) Coordinated Deception Traffic: In this case, all network nodes want to cooperatively transmit the deception traffic to maximize the impact of anti-inference. However, deception traffic unavoidably degrades the network performance since redundant data is transmitted in the network. Hence, we must limit the total amount of the deception traffic in the network. In particular, we limit the average rate of deception traffic to be $\|\mathbf{J}\|_1/n = m_c$, and limit individual rate to be $J_i \leq \sigma_c$, where m_c and σ_c are some positive constants. Given these limited costs, we present the following theorem to show the impact of coordinated deception traffic. Note that in order to be more consistent and informative, we place all proofs of theorems in Section IV.

Theorem 1: Under the coordinated deception traffic strategy $\mathcal{J}_{D}$ that all nodes transmit deception traffic with rate vector $\mathbf{J} = \{J_i\}_{i\in[1,L]}$ unknown to the attacker, and satisfying average rate constraint $\|\mathbf{J}\|_1/n = m_c$ and individual rate constraint $J_i \leq \sigma_c$ for some positive constants m_c and σ_c, the genie bound of network inference is given by

$$\Theta\left(\frac{m_c n^2}{g(n)(n+Fg(n))}\right) \leq \mathcal{G}(\mathbf{x}_g|\mathcal{J}_{D}) \leq \Theta\left(\frac{\sigma_c^2 n}{g(n)}\right), \tag{5}$$

where n is the number of nodes, F is the number of end-to-end flows, and $g(n)$ is defined in Model 1.

2) Random Deception Traffic: In this case, each node independently transmits random deception traffic. The random deception traffic rate of each node has a bounded mean m_{J} and a bounded variance σ_{J}^2. We present the following theorem to show the impact of random deception traffic.

Theorem 2: Under the random deception traffic strategy $\mathcal{J}_{R}$ that each node independently transmits deception traffic at a rate that follows an arbitrary distribution with bounded mean m_{J} and variance σ_{J}^2, if the attacker is unaware of the value of m_{J}, the genie bound of network inference satisfies

$$\Theta\left(\frac{\sigma_{J}^2 F}{g(n)}\right) \leq \mathcal{G}(\mathbf{x}_g|\mathcal{J}_{R}) \leq \Theta\left(\frac{(m_{J}^2+\sigma_{J}^2)n}{g(n)}\right). \tag{6}$$

If the attacker knows m_{J}, the genie bound becomes

$$\mathcal{G}(\mathbf{x}_g|\mathcal{J}_{R}) = \Theta\left(\frac{\sigma_{J}^2 F}{g(n)}\right). \tag{7}$$

3) Observations and Comparisons: Theorems 1 and 2 show that if we are allowed to transmit a limited amount of deception traffic (that leads to limited throughput degradation) to affect network inference, the genie bound (i.e., the error under the theoretically optimal inference method) is at most on the order of $n/g(n)$ for either the random or any coordinated transmission strategies (see the upper bounds in (5) and (6)). In practice, compared with the simple random strategy, the very best coordinated strategy may require much cooperation among nodes, yet still causing the inference error on the same order of magnitude. Consequently, the random transmission strategy can be desirable for simple and efficient deployment of network anti-inference.

Remark 3: Theorem 2 states that if the mean rate of the random transmission strategy is known to the attacker, the induced error (7) is only the lower bound in (6). This means that the practical design of a random transmission strategy should always attempt to hide the mean rate from the attacker.

Remark 4: It is also observed that the genie bounds in Theorems 1 and 2 increase under routing protocols that yield shorter paths (i.e., $g(n)$ becomes smaller), indicating that a shorter routing path helps cause more inference errors. Intuitively, if a network flow is routed over a longer path, it provides more statistics for observation, and therefore can be better inferred by the attacker. Thus, the shortest-path routing in fact helps anti-inference under the deception traffic strategy.

Remark 5: It is worth emphasizing that the results given in Theorems 1 and 2 are represented by the genie bound instead of the actual estimation error due to a particular inference algorithm. In practice, an inference algorithm always results in a performance gap between the genie bound and the real inference error.

D. Routing Changing

Next, we investigate how the routing changing strategy affects network inference. As we have seen previously, it is relatively easy to formulate the effect of deception traffic as an additive noise in (4). Now, we need to analyze how effective a routing changing strategy can be. Such a strategy will cause data to be forwarded on a new routing path, leading to a routing matrix $\mathbf{B}$ that is not equal to the original routing matrix $\mathbf{A}$, i.e., $\mathbf{B} \neq \mathbf{A}$. Our objective is to compute the genie bound of the network inference due to the routing matrix mismatch.

In addition, we also need to measure the cost of routing changing. Thanks to Model 1, we denote by $g(n)$ and $h(n)$ the average numbers of routing hops due to the original and new routing protocols, respectively. We assume that $h(n) \geq g(n)$. This means that the cost of the routing changing strategy can be immediately quantified by $h(n)/g(n) \geq 1$. For such a routing changing strategy, we present the following result.

Theorem 3: Under the routing changing strategy $\mathcal{R}$ in which (i) the original routing matrix $\mathbf{A}$ is changed to an independent matrix $\mathbf{B}$ unknown to the attacker, and (ii) the number of hops $g(n)$ is changed to $h(n) \geq g(n)$ that satisfies Model 1, the genie bound of network anti-inference satisfies

$$\Theta\left(\frac{(m_{L}^2+\sigma_{L}^2)F^2h(n)^2}{(n+Fg(n))g(n)}\right) \leq \mathcal{G}(\mathbf{x}_g|\mathcal{R}) \leq \Theta\left(\frac{(m_{L}^2+\sigma_{L}^2)F^2h(n)}{g(n)}\right), \tag{8}$$

where m_{L} and σ_{L}^2 are the mean and variance of each legitimate flow's rate, respectively.

Remark 6: Theorem 3 provides a general impact region for any routing changing strategy. We note that in order to limit the cost of such a strategy, the new routing path $h(n)$ should be on the same order of $g(n)$, i.e., $h(n)/g(n) = \Theta(1)$. This means that a message should be routed over the new path longer than the original one by only a constant order of magnitude, leading to (roughly speaking) a constant increase in the end-to-end delay. If the cost is allowed to be of higher order than $\Theta(1)$, we conclude by looking at the lower bound in (8) that increasing the new routing path length $h(n)$ can incur much more inference errors.

E. Combining Deception Traffic and Routing Changing

Finally, we evaluate the impact of the strategy that combines deception traffic and routing changing.

Theorem 4: If each node proactively transmits random deception traffic and changes its routing path, the induced genie bound is on the highest order between the individual bounds of the two strategies.

Remark 7: From Theorem 4, we know that when the deception traffic and routing changing strategies are combined, the impact mainly depends on whichever strategy that can lead to more impact. However, the combined strategy indeed leads to substantially more costs in the sense that it requires both transmitting deception traffic and changing routing. Hence, the combined strategy can be avoided to limit the cost when one proactive strategy (deception traffic or routing changing) is known to be better than the other.

F. Examples and Discussions

We have analyzed the impact of each proactive strategy with a bounded cost on network inference. As we observe, errors of network inference caused by proactive strategies mainly depend on the number of flows F, the number of nodes n, and the routing paths $g(n)$, $h(n)$. It is not intuitive to directly compare their impacts to see which one is better than the other.

We use examples to compare the impacts of proactive strategies. In particular, we consider the scenario in which each node is communicating with a limited number of other nodes (i.e., $F=\Theta(n)$), and operates under the best routing. It can be verified that $g(n) \geq \Theta(\sqrt{n})$ under any routing protocol and we choose $g(n) = \Theta(\sqrt{n})$ as an example of the best routing. If the routing changing strategy is used, each node chooses a different path but on the same order of $g(n)$ (i.e., $h(n)=\Theta(\sqrt{n})$) such that the delay degradation is bounded.

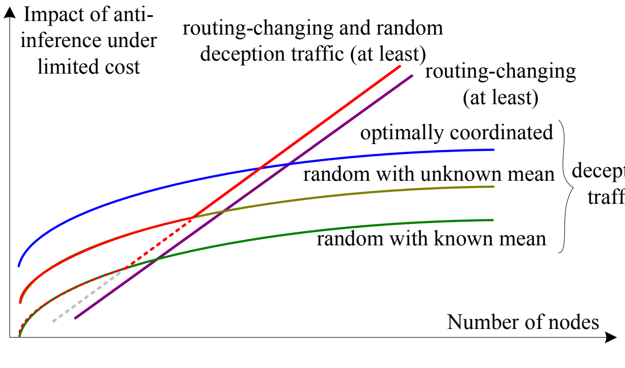


Fig. 3. The impact (estimation error caused by anti-inference) for different proactive strategies with bounded costs.

Fig. 3 illustrates the impacts (i.e., genie bounds) of different strategies (according to the theoretical predictions). We can observe that all three deception traffic strategies are on the same order of $\sqrt{n}$, Therefore, hiding the mean value of the deception traffic rate from the attacker or coordinating all nodes in the optimal way only leads to constant improvement over the simplest strategy that transmits deception traffic independently at each node.

We also see from Fig. 3 that the routing changing strategy at least leads to inference errors on the order of n, which indicates that in this typical network scenario, changing routing is a substantially better strategy than transmitting deception traffic (both under limited costs). In addition, the combined strategy is better than others, but is still on the order of n.

It is worth noting that the genie bound is an absolute metric. Our objective is to find out which anti-inference strategy is better under what conditions by comparing different genie bounds. For example in Fig. 3, by comparing the genie bounds of the deception traffic strategies, we can know that the error is only increased by a constant magnitude when we optimally coordinate the transmissions of deception traffic across the network.

G. Applications

We note that the objective of this paper is not focused on designing a detailed deception protocol to fool network inference, but on the fundamental perspective on the impacts of exploring network dynamics (in terms of transmitting deception traffic or changing routing) with limited costs to offer more security for network nodes. Therefore, the applications of our results include

- showing the impact region of a proactive strategy with a limited cost for a given network scenario,
- providing the performance benchmark and guidance for anti-inference protocol design,
- offering a counterpart strategy that can further advance network inference methods.

IV. Proofs of Theorems

In this section, we provide detailed proofs of Theorems 1, 2, 3, and 4, respectively.

A. Impact of Transmitting Deception Traffic

We first prove Theorem 1 that reveals the impact of coordinated deception traffic, then prove Theorem 2 that shows the impact of random deception traffic.

Proof of Theorem 1: Recall that given deception traffic rate vector $\mathbf{J}$, the under-determined system for network inference is

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{J}, \tag{9}$$

where $\mathbf{x} \in \mathbb{R}^{(n(n-1)/2)\times 1}$ is the rate vector for all possible end-to-end flows, $\mathbf{y} \in \mathbb{R}^{L\times 1}$ is the observation vector for all links, and $\mathbf{A} \in \mathbb{R}^{L\times(n(n-1)/2)}$ is the routing matrix. By only considering the genie bound, we can re-write (9) as

$$\mathbf{y} = \mathbf{A}_g\mathbf{x}_g + \mathbf{J}, \tag{10}$$

where $\mathbf{x}_g \in \mathbb{R}^{F\times 1}$ is the flow rate vector for all existing end-to-end flows and $\mathbf{A}_g \in \mathbb{R}^{L\times F}$ is the routing matrix for existing end-to-end flows. The minimum mean squared error estimate of x_g can be obtained by performing the least squares estimation as

$$\hat{\mathbf{x}}_g = \underset{\mathbf{x}_g\in\mathbb{R}^{L\times 1}}{\arg\min}\|\mathbf{y}-\mathbf{A}_g\mathbf{x}_g\|_2^2 = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y}. \tag{11}$$

It follows from (10) and (11) that the genie bound is[1]

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2=\mathbb{E}\left(\|(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{J}\|_2^2\right)=\|\mathbf{G}\mathbf{J}\|_2^2, \quad (12)$$

where $\mathbf{G}=(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T$. It follows from Lemma A1 in Appendix that

$$\lambda_{\min}\big(\mathbf{G}^T\mathbf{G}\big)\,\|\mathbf{J}\|_2^2\leq\|\mathbf{G}\mathbf{J}\|_2^2\leq\lambda_{\max}\big(\mathbf{G}^T\mathbf{G}\big)\|\mathbf{J}\|_2^2, \quad (13)$$

Then, according to Lemmas A2 and A3 in Appendix, from (13), we can have that with high probability,

$$\lambda_{\min}\big(\mathbf{G}^T\mathbf{G}\big)\Theta(m_cn)\leq\|\mathbf{G}\mathbf{J}\|_2^2\leq\lambda_{\max}\big(\mathbf{G}^T\mathbf{G}\big)\Theta(\sigma_c^2n), \quad (14)$$

To derive the maximum and minimum eigenvalues of $\mathbf{G}^T\mathbf{G}$, we look at the singular value decomposition of $\mathbf{A}_g$, which is written as $\mathbf{A}_g=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{\Sigma}$ is a rectangular diagonal matrix with non-zero values $\{\sqrt{\lambda_i(\mathbf{A}_g^T\mathbf{A}_g)}\}_{i\in[1,F]}$, $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. We obtain $\mathbf{G}=(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T=\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$, where $\mathbf{\Sigma}^{-1}$ is obtained by taking the reciprocal of each non-zero element on the diagonal, leaving the zeros in place in $\mathbf{\Sigma}$. Accordingly, $\mathbf{G}^T\mathbf{G}=\mathbf{U}(\mathbf{\Sigma}^{-1})^2\mathbf{U}^T$, which means that $\lambda_{\max}\left(\mathbf{G}^T\mathbf{G}\right)=\lambda_{\min}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)$ and $\lambda_{\min}\left(\mathbf{G}^T\mathbf{G}\right)=\lambda_{\max}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)$. Thus,

$$\lambda_{\max}^{-1}\big(\mathbf{A}_g^T\mathbf{A}_g\big)\Theta(m_cn)\leq\|\mathbf{G}\mathbf{J}\|_2^2\leq\lambda_{\min}^{-1}\big(\mathbf{A}_g^T\mathbf{A}_g\big)\Theta(\sigma_c^2n). \quad (15)$$

Finally, it follows from Lemma A4 and (15) that

$$\Theta\left(\frac{m_cn^2}{g(n)(n+Fg(n))}\right)\leq\|\mathbf{G}\mathbf{J}\|_2^2\leq\Theta\left(\frac{\sigma_c^2n}{g(n)}\right), \quad (16)$$

which completes the proof. □

Proof of Theorem 2 (Part I): We first consider the case that the attacker knows the value of $m_{_J}$. In this case, $\mathbf{y}$ is the link observation vector affected by the deception traffic with mean rate $m_{_J}$. This indicates that the least squares estimate of x_g can be obtained by first subtracting each observed rate by $m_{_J}$ then performing the least squares estimation as

$$\begin{aligned}\hat{\mathbf{x}}_g &= \operatorname*{arg\,min}_{\mathbf{x}_g\in\mathbb{R}^{L\times 1}}\|\mathbf{y}-\mathbf{m}_{_J}-\mathbf{A}_g\mathbf{x}_g\|_2^2\\ &= (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T(\mathbf{y}-\mathbf{m})=\mathbf{G}(\mathbf{y}-\mathbf{m}), \quad (17)\end{aligned}$$

where $\mathbf{G}=(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T$, and $\mathbf{m}_{_J}=[m_{_J},m_{_J},\cdots,m_{_J}]^T\in\mathbb{R}^{L\times 1}$. It follows from (10) and (17) that the genie bound is

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2=\mathbb{E}\|\mathbf{G}(\mathbf{J}-\mathbf{m})\|_2^2 \quad (18)\\ &=\mathbb{E}\big(\mathbf{tr}\left\{\mathbf{G}\mathbf{C}(\mathbf{J}-\mathbf{m})\mathbf{G}^T\right\}|\mathbf{A}_g\big),\end{aligned}$$

where $\mathbf{C}(\mathbf{J}-\mathbf{m})=\mathbb{E}((\mathbf{J}-\mathbf{m})(\mathbf{J}-\mathbf{m})^T)$ is the covariance matrix of $\mathbf{J}-\mathbf{m}$. Because each node transmits deception traffic independently, $\mathbf{C}(\mathbf{J}-\mathbf{m})=\sigma_{_J}^2\mathbf{I}_L$, where $\mathbf{I}_L$ denotes the $L\times L$ identity matrix. Accordingly, we have

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 &= \mathbb{E}\left(\mathbf{tr}\left\{\sigma_{_J}^2(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\right\}\right)\\ &= \sigma_{_J}^2\mathbb{E}\left(\sum_{i=1}^F\lambda_i^{-1}(\mathbf{A}_g^T\mathbf{A}_g)\right), \quad (19)\end{aligned}$$

where $\lambda_i(\mathbf{A}_g^T\mathbf{A}_g)$ is the i-th eigenvalue of matrix $\mathbf{A}_g^T\mathbf{A}_g$.

According to Lemma A5, there exists a function

$$c=\Theta\left(\sqrt{\frac{n}{g(n)}}\right) \quad (20)$$

such that each element in $c\mathbf{A}_g$ has finite mean and variance 1. We then re-write (19) as

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 &=\sigma_{_J}^2c^2\mathbb{E}\left(\frac{\sum_{i=1}^F\lambda_i^{-1}\big(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\big)}{L}\right) \quad (21)\\ &\geq\sigma_{_J}^2c^2\mathbb{E}\left(\frac{\frac{F^2}{L}}{\sum_{i=1}^F\lambda_i\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)}\right) \quad (22)\\ &\geq\sigma_{_J}^2c^2\frac{\frac{F}{L}}{\mathbb{E}\left(\frac{\sum_{i=1}^F\lambda_i(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g))}{F}\right)}, \quad (23)\end{aligned}$$

where (22) follows from the Cauchy-Schwarz inequality, and (23) follows from the property of expectation.

According to Theorem of Universality for Bulk Convergence [21]–[23], as $L\to\infty$, the probability measure for $\frac{1}{F}\sum_{i=1}^F\lambda_i\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)$ converges in distribution to the Marchenko-Pastur law, indicating that

$$\mathbb{E}\left(\frac{\sum_{i=1}^F\lambda_i\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)}{F}\right)=\Theta\left(1\right). \quad (24)$$

Inserting (24) into (23) and using the fact that $L=\Theta(n)$ with high probability in Lemma A2 lead to

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2\geq\sigma_{_J}^2c^2\Theta\left(\frac{F}{n}\right). \quad (25)$$

Inserting (20) into (25) yields

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2\geq\Theta\left(\frac{\sigma_{_J}^2F}{g(n)}\right). \quad (26)$$

On the other hand, starting from (21), we have

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2\leq\sigma_{_J}^2\mathbb{E}\left(\sum_{i=1}^F\lambda_{\min}^{-1}\big(\mathbf{A}_g^T\mathbf{A}_g\big)\right). \quad (27)$$

Then, it follows from Lemma A4 that

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2\leq\Theta\left(\frac{\sigma_{_J}^2F}{g(n)}\right), \quad (28)$$

which is combined with (26) to finish the first part. □

Proof of Theorem 2 (Part II): We then consider the case that the attacker does not know the value of $m_{_J}$. In this case, the attacker cannot subtract $m_{_J}$ from each observed rate. Thus,

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 &= \mathbb{E}\|\mathbf{G}\mathbf{J}\|_2^2\\ &= \mathbb{E}\|\mathbf{G}(\mathbf{J}-\mathbf{m})+\mathbf{G}\mathbf{m}\|_2^2\\ &\geq \mathbb{E}\|\mathbf{G}(\mathbf{J}-\mathbf{m})\|_2^2=\Theta\left(\frac{\sigma_{_J}^2F}{g(n)}\right), \quad (29)\end{aligned}$$

in which the last equality follows from (26). On the other hand, we have

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 &= \mathbb{E}\|\mathbf{G}\mathbf{J}\|_2^2\leq\mathbb{E}(\lambda_{\max}(\mathbf{G}^T\mathbf{G}))\mathbb{E}\|\mathbf{J}\|_2^2\\ &= \mathbb{E}\left(\frac{1}{\lambda_{\min}(\mathbf{A}_g^T\mathbf{A}_g)}\right)\mathbb{E}\|\mathbf{J}\|_2^2\\ &= \Theta\left(\frac{(m_{_J}^2+\sigma_{_J}^2)n}{g(n)}\right), \quad (30)\end{aligned}$$

[1] In all proofs, we slightly abuse the notation and write the genie bound under a strategy $\mathcal{S}$ as $\mathcal{G}(\mathbf{x}_g|\mathcal{S})=\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2$ instead of $\mathbb{E}(\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2|\mathcal{S})$ for the sake of simplicity.

where the last inequality follows from Lemma A1, and the last equality follows from Lemmas A2 and A4. Combining (29) and (30) finishes the second part of the proof. □

B. Impact of Changing Routing

In this subsection, we prove Theorem 3 to show the impact of routing changing on network inference.

Proof of Theorem 3: Under the proactive routing changing strategy, the attacker will have a mismatched routing matrix $\mathbf{A}_g$ (compared to the full routing matrix $\mathbf{A}$) for inference instead of the true matrix $\mathbf{B}_g$ (compared to the full routing matrix $\mathbf{B}$). Thus, the genie-assisted least squares solution for the attacker becomes $\hat{\mathbf{x}}_g = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y}$. Then, the genie bound due to using a mismatched routing matrix $\mathbf{A}_g$ to solve the true linear system $\mathbf{y} = \mathbf{B}_g\mathbf{x}_g$ can be written as

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 &= \mathbb{E}\|(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y} - \mathbf{x}_g\|_2^2 \\ &= \mathbb{E}\|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2, \qquad (31)\end{aligned}$$

where $\mathbf{G} = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T$. It follows from Lemma A1 in Appendix that

$$\begin{aligned}\lambda_{\min}(\mathbf{G}^T\mathbf{G})\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \\ \lambda_{\max}(\mathbf{G}^T\mathbf{G})\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2.\end{aligned}$$

Since $\lambda_{\max}\left(\mathbf{G}^T\mathbf{G}\right) = \lambda_{\min}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)$ and $\lambda_{\min}\left(\mathbf{G}^T\mathbf{G}\right) = \lambda_{\max}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)$, we have

$$\frac{\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2}{\lambda_{\max}\left(\mathbf{A}_g^T\mathbf{A}_g\right)} \leq \|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \frac{\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2}{\lambda_{\min}\left(\mathbf{A}_g^T\mathbf{A}_g\right)}.$$

According to Lemma A4 in Appendix, we can further have

$$\frac{\mathbb{E}\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2}{g(n) + \frac{Fg(n)^2}{n}} \leq \mathbb{E}\|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \frac{\mathbb{E}\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2}{\Theta(g(n))}. \qquad (32)$$

Next, we proceed to derive $\mathbb{E}\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2$ in (32). Denote entries in $(\mathbf{B}_g - \mathbf{A}_g)$ as $\{s_{l,f}\}_{l\in[1,L],f\in[1,F]}$ and entries in $\mathbf{x}_g$ as $\{x_f\}_{f\in[1,F]}$. According to Lemma 1, we can write $s_{l,f}$ as

$$s_{l,f} = \begin{cases} 1 & \text{with probability } \Theta(\frac{h(n)}{n})(1-\Theta(\frac{g(n)}{n})) \\ -1 & \text{with probability } \Theta(\frac{g(n)}{n})(1-\Theta(\frac{h(n)}{n})) \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbb{E}\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2$ can be represented as

$$\begin{aligned}\mathbb{E}\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 &= \mathbb{E}\left(\sum_{l=1}^{L}\left(\sum_{f=1}^{F}s_{l,f}x_f\right)^2\right) \\ &= \Theta(n)\mathbb{E}\left(\left(\sum_{f=1}^{F}s_{l,f}x_f\right)^2\right) \\ &\geq \Theta(n)\left(\mathbb{E}\left(\sum_{f=1}^{F}s_{l,f}x_f\right)\right)^2 \\ &= \Theta(n)F^2\Theta\left(\frac{h(n)^2}{n^2}\right)\mathbb{E}(x_f)^2 \\ &= \Theta\left(\frac{F^2(m_L^2+\sigma_L^2)h(n)^2}{n}\right). \qquad (33)\end{aligned}$$

On the other hand, it follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}\mathbb{E}\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 &= \Theta(n)\mathbb{E}\left(\sum_{f=1}^{F}s_{l,f}x_f\right)^2 \\ &\leq \Theta(n)\mathbb{E}\left(\sum_{f=1}^{F}s_{l,f}^2\right)\mathbb{E}\left(\sum_{f=1}^{F}x_f^2\right) \\ &= F^2\Theta(h(n))\left(m_L^2+\sigma_L^2\right). \qquad (34)\end{aligned}$$

Combining (32), (33), and (34) completes the proof. □

C. Impact of Combination of Transmitting Deception Traffic and Changing Routing

After obtaining Theorems 2 and 3, we are ready to investigate the impact of the combined strategy on network inference.

Proof of Theorem 4: Under both routing changing and deception traffic strategies, the attacker will have a mismatched routing matrix $\mathbf{A}_g$ (compared to the full routing matrix $\mathbf{A}$) for inference instead of the true matrix $\mathbf{B}_g$ (compared to the full routing matrix $\mathbf{B}$). At the same time, all nodes transmit deception traffic with rate vector $\mathbf{J}$. Thus, the genie bound is

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 &= \mathbb{E}\|\mathbf{G}\mathbf{y} - \mathbf{x}_g\|_2^2 \\ &= \mathbb{E}\|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x} + \mathbf{G}\mathbf{J}\|_2^2 \\ &= \mathbb{E}\|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}\|_2^2 + \mathbb{E}\|\mathbf{G}\mathbf{J}\|_2^2 + \\ &\qquad 2\mathbb{E}\left(\mathbf{J}^T\mathbf{G}^T\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}\right), \qquad (35)\end{aligned}$$

where $\mathbf{G} = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T$. It is straightforward to observe that $\mathbb{E}\|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}\|_2^2$ is the genie bound of routing changing, $\mathbb{E}\|\mathbf{G}\mathbf{J}\|_2^2$ is the genie bound of transmitting random traffic, and $\mathbb{E}\left(\mathbf{J}^T\mathbf{G}^T\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}\right)$ is of no higher order than them. Consequently, we conclude that the genie bound of the combined strategy is on the highest order between the bounds of two individual strategies. □

V. Simulation Results

In this section, we use numerical simulations to measure the impacts of proactive strategies. Our objective is to use theoretical results to explain the observations in numerical simulations with proactive strategies. We first introduce the network and strategy setups, then present the results.

A. Setups

1) Network Setups: We randomly distribute $n \in [50, 1200]$ nodes over network region $[0, \sqrt{n/\lambda}]^2$ where node density $\lambda = 4$, and the communication range of each node is $r = 1$. Apparently, if the network has a low node density or short communication range, it is difficult to find more paths between a source and a destination for routing changing. In simulations, we find $\lambda = 4$ and $r = 1$ yield a number of available alternative paths between two nodes.

2) Deception Traffic Strategy: Each end-to-end flow has a random rate uniformly distributed in $[0, 0.2]$. When the deception traffic strategy is enabled, each node will independently transmit deception traffic on each link with a random rate uniformly distributed in $[0, 0.1]$. We choose a deception traffic rate comparable to the flow rate to make the results evident.

3) Routing Changing Strategy: We set that the routing protocol yields $g(n) = \Theta(\sqrt{n})$. When the routing changing strategy is enabled, we use an alternative routing protocol unknown to the attacker. Specifically, for a node transmitting packets to a destination, we first select possible paths with length smaller than $2\sqrt{n}$ to form a path set. The value of $2\sqrt{n}$ is the threshold of the length of a path in simulations such that longer paths should not be chosen to severely delay the packet delivery. Then, the node will transmit each packet over a path randomly selected from this path set. The average path length over all nodes is chosen and measured as $h(n) = 1.3g(n)$.

4) Network Inference Scenarios: As we can see in the theoretical results, the genie bound is affected by the number of flows F in the network. Thus, we consider two different network inference scenarios based on the number of flows: (i) the $F = \lfloor\sqrt{n}\rfloor$ scenario in which there are limited (i.e., $\lfloor\sqrt{n}\rfloor$) flows between randomly chosen nodes in the network, and (ii) the $F = n$ scenario in which every node has a flow associated with another randomly chosen node.

5) Measuring the Genie Bound: In simulations, the genie bound is computed as follows. Step 1: the genie always knows which entries in $\mathbf{x}$ have value zero (i.e., no end-to-end flow) in (1), the genie removes these entries from $\mathbf{x}$ in (1) and accordingly establishes a new equation $\mathbf{y} = \mathbf{A}_g\mathbf{x}_g$, where $\mathbf{x}_g$ is formed by removing zero entries in $\mathbf{x}$, $\mathbf{y}$ is the observed link rate vector, and $\mathbf{A}_g$ is the routing matrix for existing end-to-end flows. Step 2: the least squares solution to the new equation is obtained as $\hat{\mathbf{x}}_g = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y}$, then the genie bound is measured as $\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 = \mathbb{E}\|(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y} - \mathbf{x}_g\|_2^2$. In this bound, only the genie knows $\mathbf{A}_g$ and $\mathbf{x}_g$. Therefore, this bound serves as the lower error bound for all practical methods that an attacker can use. In our simulations, the genie bound for each anti-inference strategy is measured over 100 simulation runs with random wireless network topologies.

B. The $F = \lfloor\sqrt{n}\rfloor$ Scenario

We compute from (6) and (7) that the genie bounds for deception traffic is $\Theta(1) \leq \mathcal{G}(\mathbf{x}_g|\mathcal{J}_{R}) \leq \Theta(\sqrt{n})$ and $\Theta(1)$ if the attacker knows and does not know the mean deception traffic rate, respectively. We also obtain from (8) that the genie bound of routing changing is $\Theta(\sqrt{n}) \leq \mathcal{G}(\mathbf{x}_g|\mathcal{R}) \leq \Theta(n)$.

Fig. 4 shows the measured genie bounds under different proactive strategies as a function of the number of nodes n. We can observe from Fig. 4 that all the measured genie bounds are sub-linear. For the deception traffic strategy shown in Fig. 4(a), if the mean deception traffic rate is known to the attacker, the measured genie bound becomes independent of n and remains as a small constant around 0.0012; otherwise, it is increasing approximately on the order of $\sqrt{n}$. As shown in Fig. 4(a), the measured genie bounds with unknown mean are curve-fitted by a closest square-root function. We also use square-root functions to curve-fit the measured genie bounds of the routing changing and combined strategies in Fig. 4(b). It is worth noting that routing changing induces higher genie bounds (i.e., more errors) for network inference. For example, when the network has 1000 nodes, Fig. 4(a) shows that the genie bound of deception traffic with unknown mean is 0.045, which is much smaller than that of routing change shown as 0.196 in Fig. 4(b). We can also see from Fig. 4(b) that the combined strategy leads to the highest genie bound, yet still on the order of $\sqrt{n}$.

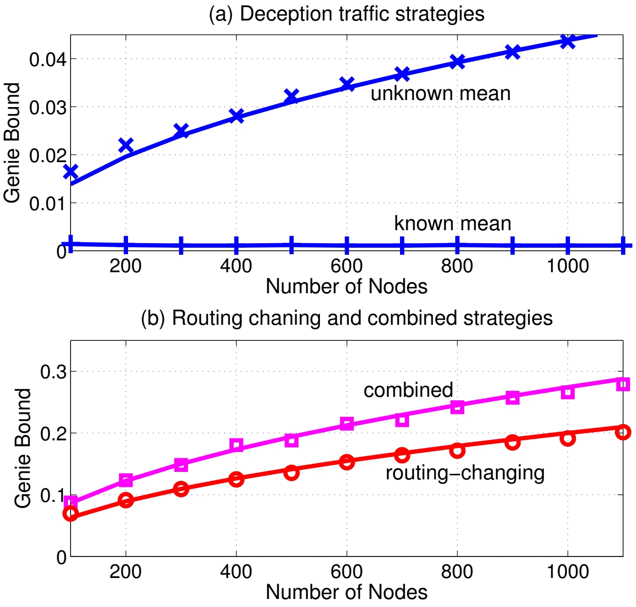


Fig. 4. Measured genie bounds under different strategies when $F = \lfloor\sqrt{n}\rfloor$.

C. The $F = n$ Scenario

When $F = n$, we compute from the theoretical results that the genie bound for deception traffic is always $\Theta(\sqrt{n})$, and that of routing changing satisfies $\Theta(n) \leq \mathcal{G}(\mathbf{x}_g|\mathcal{R}) \leq \Theta(n^2)$.

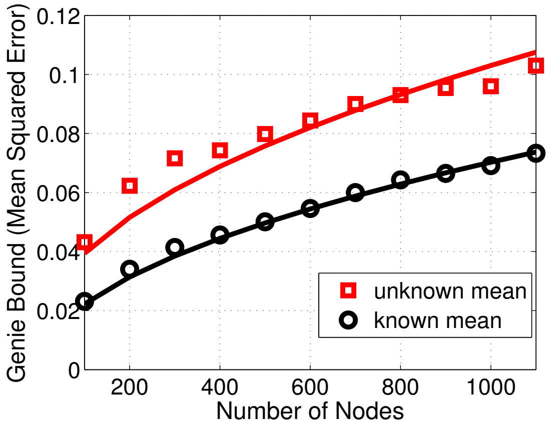


Fig. 5. Measured genie bounds under deception traffic when $F = n$.

Fig. 5 shows the measured genie bounds curve-fitted by closest square-root functions under the deception traffic strategies. We can see that although the genie bound is still higher when the attacker does not know the mean deception traffic rate, the two bounds are both on the order of $\sqrt{n}$, which differs from Fig. 4.

Fig. 6 shows the measured genie bounds under the routing changing and combined strategies versus the number of nodes

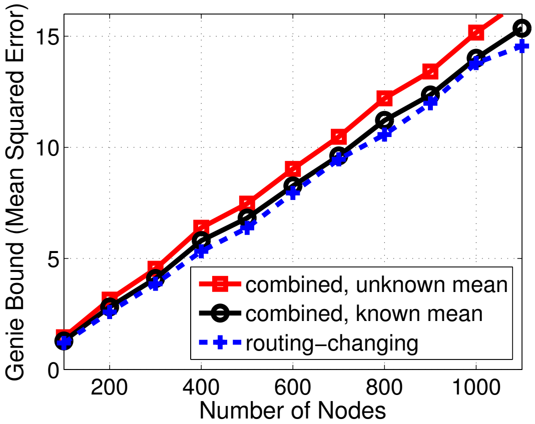


Fig. 6. Measured genie bounds under the routing changing and combined strategies in the $F = n$ scenario.

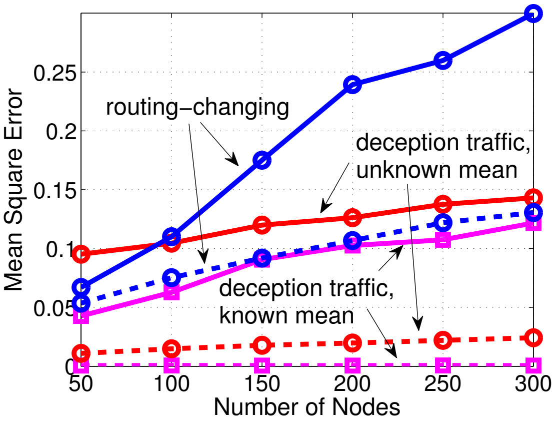


Fig. 7. Mean square error of the in-crowd algorithm under different strategies.

n. We can see that all the genie bounds increase linearly, and the gaps between the three strategies are small. Moreover, we can see that routing changing and combined strategies induce substantially more errors compared with deception traffic shown in Fig. 5.

By comparing with simulation results in Figs. 4, 5, and 6 with theoretical predictions, we conclude that the simulations validate the theoretical results on proactive strategies with bounded costs. In addition, we see that changing routing is overall a better strategy than transmitting deception traffic.

D. Impacts of Proactive Strategies on A Practical Algorithm

Our previous simulations are based on measuring the genie bound that represents the theoretically optimal performance for flow rate estimators. In the literature, network inference is generally solved as a basis pursuit denoising problem. Thus, in the following, we use the in-crowd algorithm [20], which is a fast and efficient method for solving basis pursuit denoising, to estimate flow rates in the network.

To make sure network inference is well-conditioned in the in-crowd algorithm, we evaluate the impact of anti-inference on the more sparse $F = \lfloor\sqrt{n}\rfloor$ case. Specifically, we randomly generate $F = \lfloor\sqrt{n}\rfloor$ flows in the network to make sure that there are only $\lfloor\sqrt{n}\rfloor$ non-zero entries in the flow rate vector $\mathbf{x}$ with length $\frac{n(n-1)}{2}$. The routing matrix $\mathbf{A}$ is always generated to have the full rank.

If there is no anti-inference, we find that the in-crowd algorithm performs quite well, resulting in a mean square error of around 1.4e-4 with 300 nodes in the network, which can be considered as the baseline performance for the in-crowd algorithm without anti-inference. Then, we show its mean square errors (in solid lines) under anti-inference strategies in Fig. 7. The genie bounds for these strategies are also drawn in dashed lines. From Fig. 7, we conclude that proactive strategies cause much more errors to practical algorithms (that are unavoidably non-optimal and lead to performance penalties compared with the genie bound) for network inference. Thus, our genie bound analysis can serve as the at-least disruption benchmark for the impacts of proactive strategies on network inference in practice.

VI. Related Work

Network inference or tomography has been shown as an effective way to infer end-to-end flow or link rates from network measurements [1], [5]–[8]. There are several underlying mathematical methods for network inference, including least squares estimation [2]–[4], sparsity recovery [6], [9], [17], [18], [20] and statistics based approaches [1], [5], some of which have been designed for noisy environments. For example, the in-crowd algorithm [20] used in this paper is a fast method to find an optimal solution with noise suppression, making it an effective method for wireless network inference. In this paper, we focus on the opposite of network inference, i.e., network anti-inference, which has applications in clandestine communication and secure networking scenarios.

A line of work related to our study in this paper is prevention of information flow detection (e.g., [13], [14], [19]). In this regard, an attacker can observe the packet transmissions over a path of interest to determine whether there exists an end-to-end flow on the path, which is essentially a binary inference problem. Transmitting redundant packets (also called chaff packets) has been proposed to defend against such binary inference. It was shown in [13] that when the ratio between chaff traffic and legitimate traffic is large enough, the attacker can never correctly infer the existence of an information flow on the path. However, this line of work only considered the binary inference problem and also assumed a simple communication scenario over an end-to-end path, where the only legitimate traffic is the traffic from the source to the destination over the path, and the other traffic is all considered as chaff or noise. This assumption is not always true in a network scenario where different legitimate traffic flows may cross paths with each other. Hence, the results on prevention of information flow detection cannot be directly adapted to a general scenario in network anti-inference.

Network anti-inference is a generic mechanism to make network inference inaccurate. To the best of our knowledge, network anti-inference is not well explored in the literature. Therefore, we focus on formulating and evaluating proactive strategy based network anti-inference in this paper.

VII. Conclusions

In this paper, we provided a fundamental view on network anti-inference against end-to-end flow estimation. We used the genie bounds to analyze the impacts of proactive strategies. We found that the random transmission strategy of deception traffic can achieve the impact on the same order of the best coordinated transmission strategy and the routing changing strategy is generally better than the deception traffic strategy. Our results revealed the theoretical perspective of exploring network dynamics to offer defense against network inference. Our future work includes comprehensive evaluation of realistic routing changing protocols (e.g., measuring $g(n)$ and $h(n)$) and the design of practical anti-inference systems.

References

[1] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu, "Network tomography: Recent developments," *Statistical Science*, vol. 19, pp. 499–517, 2004.

[2] T. Bu, N. Duffield, F. L. Presti, and D. Towsley, "Network tomography on general topologies," in *Proc. of ACM SIGMETRICS*, 2002.

[3] J. D. Horton and A. Lopez-Ortiz, "On the number of distributed measurement points for network tomography," in *Proc. of ACM SIGCOMM IMC*, 2003, pp. 204–209.

[4] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot, "Traffic matrices: Balancing measurements, inference and modeling," in *Proc. of ACM SIGMETRICS*, 2005.

[5] Y. E. Sagduyu, Y. Shi, A. Fanous, and J. H. Li, "An analytical framework and implementation of wireless network inference and optimization," in *Proc. of IEEE Globecom*, 2013.

[6] H. Yao, S. Jaggi, and M. Chen, "Network coding tomography for network failures," in *Proc. of IEEE INFOCOM*, 2010.

[7] A. Chen, J. Cao, and T. Bu, "Network tomography: Identifiability and fourier domain estimation," *IEEE Trans. Signal Processing*, vol. 58, pp. 6029–6039, 2010.

[8] Q. Zhao, Z. Ge, J. Wang, and J. Xu, "Robust traffic matrix estimation with imperfect information: Making use of multiple data sources," in *Proc. of ACM SIGMETRICS*, 2006, pp. 133–144.

[9] M. H. Firooz and S. Roy, "Link delay estimation via expander graphs," *IEEE Trans. Communications*, vol. 62, pp. 170–180, 2014.

[10] T. Plessea, C. Adjihb, P. Minetb, A. Laouitib, A. Plakoob, M. Badelb, P. Muhlethalerb, P. Jacquet, and J. Lecomtea, "OLSR performance measurement in a military mobile ad hoc network," *Ad Hoc Networks*, pp. 575–588, 2004.

[11] L. Ma, T. He, K. K. Leung, A. Swami, and D. Towsley, "Identifiability of link metrics based on end-to-end path measurements," in *Proc. of ACM IMC*, 2013.

[12] M. Penrose, *Random Geometric Graphs*. Oxford Univ. Press, 2003.

[13] T. He and L. Tong, "Detection of information flows," *IEEE Trans. Information Theory*, vol. 54, pp. 4925–4944, Nov. 2008.

[14] J. Kim and L. Tong, "Unsupervised and nonparametric detection of information flows," *Signal Processing*, vol. 11, Nov. 2012.

[15] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate information," *Communications on Pure and Applied Mathematics*, pp. 1207–1233, 2005.

[16] ——, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, pp. 489–509, 2006.

[17] C.-K. Yu, K.-C. Chen, and S.-M. Cheng, "Cognitive radio network tomography," *IEEE Trans. Vehicular Technology*, vol. 59, 2010.

[18] A. Krishnamurthy and A. Singh, "Robust multi-source network tomography using selective probes," in *Proc. of IEEE INFOCOM*, 2012.

[19] A. Blum, D. Song, and S. Venkataraman, "Detection of interactive stepping stones: Algorithms and confidence bounds," in *Proc. of the RAID Symposium*, 2004, pp. 258–277.

[20] P. R. Gill, A. Wang, and A. Molnar, "The in-crowd algorithm for fast basis pursuit denoising," *IEEE Trans. Signal Processing*, vol. 59, pp. 4595 – 4605, 2011.

[21] D. Chafa, "Singular values of random matrices," *Lecture Notes*, 2009.

[22] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, 2nd ed. Springer Series in Statistics, 2010.

[23] Z. D. Bai and Y. Q. Yin, "Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix," *Annals of Probability*, vol. 21, pp. 1275–1294, 1993.

[24] B. N. Parlet, *The symmetric eigenvalue problem, Classics in Applied Mathematics*. SIAM, 1998.

[25] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge Univ. Press, 2005.

[26] Z. S. Szewczak, "On Marcinkiewicz-Zygmund laws," *Journal of Mathematical Analysis and Applications*, vol. 375, pp. 738–744, 2011.

Appendix

In this appendix, we prove all lemmas used in our theoretical analysis.

Proof of Lemma 1: First, we know that the value of $a_{i,j}$ is either 0 or 1 (see the example in Fig. 1). And $a_{i,j}$ has value 1 when link i is on routing path$j$, and value 0 otherwise.

Hence, $\mathbb{P}(a_{i,j} = 1)$ denotes the probability that link i is on routing path j, i.e., $\mathbb{P}(a_{i,j} = 1) = \mathbb{P}(\text{link } i \text{ is on routing path } j)$. If there are l fixed links in the network and routing path j consists of f fixed links, we have

$$\begin{aligned}\mathbb{P}(a_{i,j} = 1|l, f) &= \frac{\binom{l-1}{f-1}}{\binom{l}{f}} \\ &= \frac{(l-1)!f!(l-f+1)!}{l!(f-1)!(l-f+1)!} \\ &= \frac{f}{l}. \end{aligned} \tag{A1}$$

Then, $\mathbb{P}(a_{i,j} = 1)$ is the expectation of $\mathbb{P}(a_{i,j} = 1|l, f)$ and can be written as

$$\begin{aligned}\mathbb{P}(a_{i,j} = 1) &= \mathbb{E}_{l,f}(\mathbb{P}(a_{i,j} = 1|l, f)) = \mathbb{E}_{l,f}\left(\frac{f}{l}\right) \\ &= \mathbb{E}_f\left(\mathbb{E}_l\left(\frac{f}{l} \mid f\right)\right). \end{aligned} \tag{A2}$$

To further proceed, we expand $\mathbb{E}_l\left(\frac{f}{l} \mid f\right)$ in (A2) to a Taylor series around $\mathbb{E}_l(l|f)$ as

$$\mathbb{E}_l\left(\frac{f}{l} \mid f\right) = \frac{f}{\mathbb{E}_l(l|f)} + f\, O\left(\frac{\mathsf{Var}_l(l|f)}{\mathbb{E}_l^3(l|f)}\right), \tag{A3}$$

where $\mathsf{Var}_f(l|f)$ is the variance of l conditioned on f. As l is the number of links in the network with n nodes, it follows the Poisson distribution according to (A6) in Lemma A2. Consequently, $\mathbb{E}_f(l|f) = \Theta(n)$ and $\mathsf{Var}_f(l|f) = \Theta(n)$. We can then write (A3) as

$$\mathbb{E}_l\left(\frac{f}{l} \mid f\right) = \frac{f}{\Theta(n)} + f\, O\left(\frac{1}{\Theta(n^2)}\right) = \frac{f}{\Theta(n)}. \tag{A4}$$

Inserting (A4) into (A2) yields

$$\mathbb{P}(a_{i,j} = 1) = \frac{\mathbb{E}(f)}{\Theta(n)}. \tag{A5}$$

In addition, f denotes the number of links on a routing path, and $\mathbb{E}(f) = g(n)$ under Model 1. Therefore, inserting $\mathbb{E}(f) = g(n)$ into (A5) completes the proof. □

Lemma A1: For a matrix $\mathbf{X}$, it always holds that $\lambda_{\min}\left(\mathbf{X}^T\mathbf{X}\right)\|\mathbf{a}\|_2^2 \leq \|\mathbf{X}\mathbf{a}\|_2^2 \leq \lambda_{\max}\left(\mathbf{X}^T\mathbf{X}\right)\|\mathbf{a}\|_2^2$ for any arbitrary vector $\mathbf{a}$.

Proof: We can observe

$$\begin{aligned}\|\mathbf{X}\mathbf{a}\|_2^2 &= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a} = \frac{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a}}{\mathbf{a}^T\mathbf{a}}\mathbf{a}^T\mathbf{a}\\ &= \frac{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a}}{\mathbf{a}^T\mathbf{a}}||\mathbf{a}||_2^2,\end{aligned}$$

where $\frac{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a}}{\mathbf{a}^T\mathbf{a}}$ is called the Rayleigh quotient [24] with maximum $\lambda_{\max}\left(\mathbf{X}^T\mathbf{X}\right)$ and minimum $\lambda_{\min}\left(\mathbf{X}^T\mathbf{X}\right)$. This finishes the proof. □

Lemma A2: In the network with n nodes and density λ over region $[0, \sqrt{\frac{n}{\lambda}}]^2$, the number of link L is on the order of n with high probability, i.e.,

$$\mathbb{P}(L = \Theta(n)) = 1 - \Theta\left(e^{-\Theta(1)n}\right).$$

Proof: The area of the region $[0, \sqrt{\frac{n}{\lambda}}]^2$ is $\frac{n}{\lambda}$. For node $i \in [1, n]$, its number of neighbors l_i follows the Poisson distribution with parameter $\pi r^2\lambda - 1$, where r is the wireless transmission range. As a result, the total number of links L in the network can be written as

$$L = \sum_{i=1}^{n} l_i \ \sim \ \text{Poisson}\ \big(n(\pi r^2\lambda - 1)/2\big). \tag{A6}$$

Thus, for some small positive constant $c_1 < \pi r^2\lambda - 1$, it follows from a Chernoff bound argument in [25] that

$$\begin{aligned}\mathbb{P}(L \geq c_1 n) &\geq 1 - \frac{e^{\frac{-n(\pi r^2\lambda - 1)}{2}}(\frac{en(\pi r^2 - 1)\lambda}{2})^{c_1 n}}{(c_1 n)^{c_1 n}}\\ &= 1 - \frac{e^{-n\pi r^2\lambda}(e\pi r^2\lambda)^{c_1 n}}{c_1^{c_1 n}}\\ &= 1 - e^{-n\pi r^2\lambda}\left(\frac{e\pi r^2\lambda}{c_1}\right)^{c_1 n}\\ &= 1 - e^{\frac{-n(\pi r^2\lambda - 1)}{2}} e^{c_1 n \log\left(\frac{e(\pi r^2 - 1)\lambda}{2c_1}\right)}\\ &= 1 - \Theta(e^{-\Theta(1)n}).\end{aligned}$$

For some large positive constant $c_2 > \pi r^2\lambda - 1$, we obtain by using a similar argument that

$$\begin{aligned}\mathbb{P}(L \leq c_2 n) &\geq 1 - e^{\frac{-n(\pi r^2 - 1)\lambda}{2}} e^{c_2 n \log\left(\frac{e(\pi r^2 - 1)\lambda}{2c_2}\right)}\\ &= 1 - \Theta(e^{-\Theta(1)n}).\end{aligned}$$

Consequently, $\mathbb{P}(L = \Theta(n)) = 1 - \Theta(e^{-\Theta(1)n})$. □

Lemma A3: Given a vector $\mathbf{J} \in \mathbb{R}^{1\times L}$ and $L = \Theta(n)$ such that $\|\mathbf{J}\|_1 = nk$ for some constant $k > 0$, it satisfies that $\|\mathbf{J}\|_2^2 \geq \Theta(kn)$.
Proof: The constraint $\|\mathbf{J}\|_1/n = k$ indicates that at least $h(n) \leq \Theta(kn)$ elements in $\mathbf{J}$ have values on the order of $\Theta\left(\frac{kn}{h(n)}\right)$. Thus, $\|\mathbf{J}\|_2^2$ at least has value

$$\|\mathbf{J}\|_2^2 = \sum_{i=1}^{h(n)} \Theta\left(\frac{k^2n^2}{h(n)^2}\right) = \Theta\left(\frac{k^2n^2}{h(n)}\right) \geq \Theta(kn),$$

which finishes the proof. □

Lemma A4: For a random matrix $\mathbf{X} \in \mathbb{R}^{L\times F}$ with entry $X_{i,j}$ ($1 \leq i \leq L$ and $1 \leq j \leq F$) having value 0 or 1 and satisfying $\mathbb{E}(X_{i,j}) = g(n)/n$ for some function $g(n) = \mathcal{O}(n)$ and $L = \Theta(n)$, if $F \to \infty$ with $\lim_{L\to\infty} F/L < \infty$, then the following holds asymptotically almost surely.

1) The minimum eigenvalue satisfies

$$\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = \Theta(g(n)).$$

2) The maximum eigenvalue satisfies

$$\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leq \Theta\left(g(n) + \frac{Fg(n)^2}{n}\right).$$

Proof: 1) According to Lemma A5, there exists a function

$$c = \Theta\left(\sqrt{\frac{n}{g(n)}}\right)$$

such that the variance of each entry in $c\mathbf{X}$ is 1. Thus, we write

$$\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = \frac{L}{c^2}\lambda_{\min}(L^{-1}(c\mathbf{X})^T(c\mathbf{X})).$$

According to [21], the eigenvalue $\lambda_{\text{min}}\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)$ converges to a positive constant asymptotically almost surely. Thus, we have

$$\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = \frac{L}{c^2}\Theta(1) = \Theta(g(n))$$

with high probability.

2) Denote $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{Y} + \frac{g(n)}{n}\mathbf{Z},$$

where $\mathbf{Z}$ is an all-one matrix and $\mathbb{E}(\mathbf{Y})$ is an all-zero matrix. Then, we have

$$\mathbf{X}^T\mathbf{X} = \mathbf{Y}^T\mathbf{Y} + \frac{g(n)}{n}\mathbf{Y}^T\mathbf{Z} + \frac{g(n)}{n}\mathbf{Z}^T\mathbf{Y} + \frac{g(n)^2}{n^2}\mathbf{Z}^T\mathbf{Z}.$$

Thus, we can write the eigenvalue of $\mathbf{X}^T\mathbf{X}$ as

$$\begin{aligned}\lambda_{\max}\left(\mathbf{X}^T\mathbf{X}\right) \quad &\leq \lambda_{\max}\left(\mathbf{Y}^T\mathbf{Y}\right) + \tfrac{2g(n)\lambda_{\max}\left(\mathbf{Y}^T\mathbf{Z}\right)}{n}\\ &\quad + \tfrac{g(n)^2\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right)}{n^2}.\end{aligned} \tag{A7}$$

It follows from [21] that

$$\lambda_{\max}\left(\mathbf{Y}^T\mathbf{Y}\right) = \Theta(g(n)) \tag{A8}$$

with high probability.

Then, we take a look at $\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right)$ in (A7). As $\mathbf{Z}$ is an all-one matrix, the rank $\left(\mathbf{Z}^T\mathbf{Y}\right)$ is 1, and

$$\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right) = \mathbf{tr}\{\mathbf{Z}^T\mathbf{Y}\} = \sum_{l=1}^{L}\sum_{f=1}^{F} y_{l,f},$$

where $y_{l,f}$ is the (l, f)-th entry in $\mathbf{Y}$. It follows from the Marcinkiewicz-Zygmund strong law of large numbers [26] that, asymptotically almost surely,

$$\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right) = o((nF)^{\frac{1}{p}})$$

for any $1 \leq p < 2$. Since $F \leq L = \Theta(n)$, we have

$$\frac{g(n)}{n}\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right) = o\left(\frac{F^{\frac{1}{p}}g(n)}{n^{1-\frac{1}{p}}}\right) \leq o\left(n^{\frac{2}{p}-1}g(n)\right). \quad \text{(A9)}$$

Next, we consider $\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right)$ in (A7). Similar to $\left(\mathbf{Z}^T\mathbf{Y}\right)$, $\mathbf{Z}^T\mathbf{Z}$ is also of rank 1. Hence, we obtain

$$\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right) = \mathbf{tr}\{\mathbf{Z}^T\mathbf{Z}\} = \sum_{l=1}^{L}\sum_{f=1}^{F} 1 = \Theta(Fn). \quad \text{(A10)}$$

Inserting (A8), (A9), and (A10) into (A7) and letting $\xi = \frac{2}{p} - 1$ yield

$$\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leq \Theta\left(n^{\xi}g(n) + \frac{Fg(n)^2}{n}\right)$$

asymptotically almost surely for any arbitrarily small $\xi > 0$, which means that

$$\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leq \Theta\left(g(n) + \frac{Fg(n)^2}{n}\right).$$

□

Lemma A5: Fr a routing matrix $\mathbf{A}$ in Lemma 1, There exists a function

$$c = \Theta\left(\sqrt{\frac{n}{g(n)}}\right)$$

such that each element in $c\mathbf{A}$ has finite mean and variance 1.

Proof: It suffices to show the means and variance of each element in $c\mathbf{A}$ are both finite. For an arbitrary element $a_{i,j}$ in $\mathbf{A}$, we have $\mathbb{P}(a_{i,j} = 1) = \Theta\left(\frac{g(n)}{n}\right)$ according to Lemma 1. Define $b = ca_{i,j}$. We can write the mean as

$$\begin{aligned} \mathbb{E}(b) &= \mathbb{E}(ca_{i,j}) = c\mathbb{E}(a_{i,j}) = c\mathbb{P}(a_{i,j} = 1) \\ &= \Theta\left(\sqrt{\frac{g(n)}{n}}\right) = \Omega(1), \end{aligned} \quad \text{(A11)}$$

and the variance as

$$\begin{aligned} \mathsf{Var}(b) &= \mathbb{E}(b^2) - \mathbb{E}^2(b) = \mathbb{E}(c^2a_{i,j}^2) - \Theta(1) \\ &= c^2\mathbb{E}(a_{i,j}^2) - \Theta(1) = c^2\Theta\left(\frac{n}{g(n)}\right) - \Theta(1) \\ &= \Theta(1). \end{aligned} \quad \text{(A12)}$$

Combining (A11) and (A12) finishes the proof. □