Network Anti-Inference: A Fundamental Perspective on Proactive Strategies to Counter Flow Inference

Zhuo Lu
Department of Computer Science
University of Memphis TN 38152
Email: zhuo.lu@memphis.edu

Cliff Wang
Army Research Office
Research Triangle Park, NC 27709
Email: cliff.wang@us.army.mil

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.**

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 clandestine communication, a node that maintains several end-to-end traffic flows does not want someone else to know whom it is communicating with. Thus, it may deploy an anti-inference strategy to make sure all the flow rates from it to other nodes are inferred as zeros. In military applications, if a commanding node keeps sending commands to others, an adversary can identify its commanding role via inferring that it has end-to-end flows to others with the same data rate.

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 constant 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. We showed that combining deception traffic and routing changing cannot significantly boost the impact of anti-inference. Rather, the

This research in this paper is partly supported by ARO staff research grant W911NF-07-R-0001-05.

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. Finally, we summarize the conclusions in Section VI.

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 $\mathbf{A}$.

II. Preliminaries and Problem Statement

In this section, we introduce models and assumptions, then state the research problem of network anti-inference.

A. Network Models

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) [10]. 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

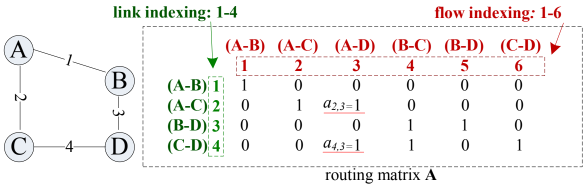
There exists an attacker attempting to use network inference to infer all end-to-end flow rates. We assume that the attacker has the strong capability of overhearing all the data transmissions (e.g., by placing eavesdroppers all over the network). The attacker is aware of the network topology; hence, given a routing protocol used in the network (e.g., shortest path routing), the attacker knows the routing path for any flow. We assume that the attacker has a perfect observation on all activities in the network. 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 $n(n-1)/2$ flows in the network. All of them 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 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], [11], [12]) that has already studied this problem. Under the condition that $\mathbf{x}$ is usually sparse (as it is less likely that everyone is communicating with everyone in practice), effective algorithms, such as $\mathcal{L}_1$-norm minimization based solutions [6], [7], [13], [14], have been developed to solve the network inference problem in many applications.

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: (i) 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. (ii) 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 [15]–[17]. 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. Main Results on Proactive Strategies against Network Inference

As the attacker can choose any method (e.g., $\mathcal{L}_1$-norm based [7], [18]) to infer the network flows, the error induced by a proactive strategy hinges on the inference method that the attacker uses. 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).

In this section, we first define the genie bound of network inference, then present the main results on the impacts of the proactive strategies against inference. We provide detailed proofs for the main results in Section IV.

A. Approach and Modeling for Anti-Inference

The genie bound is a lower bound [11], [12] of errors to solve (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. The genie bound is defined as follows.

Definition 1: Conditioned on a proactive strategy $\mathcal{S}$, the genie bound 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 are ready to analyze how proactive strategies affect the genie bound of network inference. 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. All nodes can transmit deception traffic either independently or coordinately. Then, we investigate how the routing changing strategy affects network inference. Finally, we evaluate the impact of the strategy that combines deception traffic and routing changing.

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.

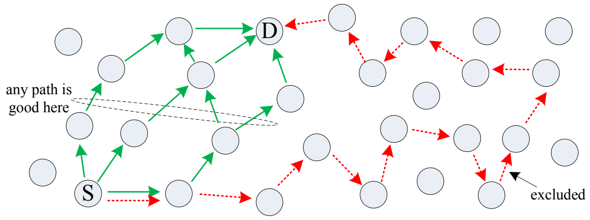


Fig. 2. Node selections in routing protocols from source S to destination D.

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).

B. Main Results

After defining the genie bound and the routing model, we summarize our main results on network anti-inference.

Theorem 1: Under the deterministic deception traffic strategy $\mathcal{J}_D$ that all nodes transmit deception traffic with deterministic 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 $J_i \le \sigma_c$ for some positive constants m_c and σ_c, the genie bound of network inference is given by

$$\Theta\left(\frac{m_c n^2}{g(n)(n+Fg(n))}\right) \le \mathcal{G}(\mathbf{x}_g|\mathcal{J}_D) \le \Theta\left(\frac{\sigma_c^2 n}{g(n)}\right), \qquad (2)$$

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

Theorem 2: Under the random deception traffic strategy $\mathcal{J}_R$ that each node independently transmits deception traffic at a random rate 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

$$\Theta\left(\sigma_J^2 F/g(n)\right) \le \mathcal{G}(\mathbf{x}_g|\mathcal{J}_R) \le \Theta\left((m_J^2+\sigma_J^2)n/g(n)\right). \qquad (3)$$

If the attacker knows m_J, the genie bound becomes

$$\mathcal{G}(\mathbf{x}_g|\mathcal{J}_R) = \Theta\left(\sigma_J^2 F/g(n)\right). \qquad (4)$$

Remark 2: 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 (2) and (3)). 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 (4) is only the lower bound in (3). 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.

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 $h(n) = \Omega(g(n))$ that satisfies Model 1, the genie bound of network anti-inference satisfies

$$\Theta\left(\frac{(m_L^2+\sigma_L^2)F^2h(n)^2}{(n+Fg(n))g(n)}\right) \le \mathcal{G}(\mathbf{x}_g|\mathcal{R}) \le \Theta\left(\frac{(m_L^2+\sigma_L^2)F^2h(n)}{g(n)}\right), \qquad (5)$$

where m_L and σ_L^2 are the mean and variance of each legitimate flow's rate.

Remark 5: 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 (5) that increasing the new routing path length $h(n)$ can incur much more inference errors.

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 6: From Theorem 4, we know that when the deception traffic and routing changing strategies are combined, the impact will not be doubled, but mainly depends on whichever strategy that can lead to more impact. However, the cost of the combined strategy is indeed doubled 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.

C. 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 $g(n) \ge \Theta(\sqrt{n})$ 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.

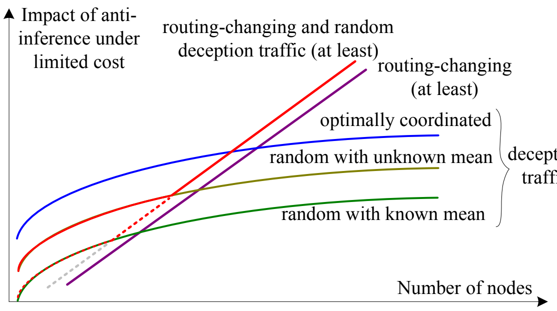


Fig. 3. The impact (estimation error caused by anti-inference) for different proactive strategies with bounded costs.

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.

D. Applications

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

IV. Network Anti-Inference Analysis

In this section, we prove all the theorems.

A. Impact of Transmitting Deception Traffic

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

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

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{J}, \tag{6}$$

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 (6) as

$$\mathbf{y} = \mathbf{A}_g\mathbf{x}_g + \mathbf{J}, \tag{7}$$

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

$$\hat{\mathbf{x}}_g = \underset{\mathbf{x}_g\in\mathbb{R}^{L\times 1}}{\arg\min}\|\mathbf{y} - \mathbf{A}_g\mathbf{x}_g\|_2^2 = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y}. \tag{8}$$

It follows from (7) and (8) that the genie bound is[1]

$$\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 = \mathbb{E}\left(\|(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{J}\|_2^2\right) = \|\mathbf{G}\mathbf{J}\|_2^2, \tag{9}$$

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

$$\lambda_{\text{min}}\left(\mathbf{G}^T\mathbf{G}\right)\|\mathbf{J}\|_2^2 \le \|\mathbf{G}\mathbf{J}\|_2^2 \le \lambda_{\text{max}}\left(\mathbf{G}^T\mathbf{G}\right)\|\mathbf{J}\|_2^2, \tag{10}$$

Then, according to Lemmas A2 and A3 in Appendix, from (10), we can have that with high probability,

$$\lambda_{\text{min}}\left(\mathbf{G}^T\mathbf{G}\right)\Theta(m_c n) \le \|\mathbf{G}\mathbf{J}\|_2^2 \le \lambda_{\text{max}}\left(\mathbf{G}^T\mathbf{G}\right)\Theta(\sigma_c^2 n), \tag{11}$$

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 $\lambda_{\text{max}}\left(\mathbf{G}^T\mathbf{G}\right) = \lambda_{\text{min}}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)$ and $\lambda_{\text{min}}\left(\mathbf{G}^T\mathbf{G}\right) = \lambda_{\text{max}}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)$. Thus,

$$\lambda_{\text{max}}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)\Theta(m_c n) \le \|\mathbf{G}\mathbf{J}\|_2^2 \le \lambda_{\text{min}}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)\Theta(\sigma_c^2 n). \tag{12}$$

Finally, it follows from Lemma A4 and (12) that

$$\Theta\left(\frac{m_c n^2}{g(n)(n + Fg(n))}\right) \le \|\mathbf{G}\mathbf{J}\|_2^2 \le \Theta\left(\frac{\sigma_c^2 n}{g(n)}\right), \tag{13}$$

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

$$\begin{aligned}\hat{\mathbf{x}}_g &= \underset{\mathbf{x}_g\in\mathbb{R}^{L\times 1}}{\arg\min}\|\mathbf{y} - \mathbf{m}_J - \mathbf{A}_g\mathbf{x}_g\|_2^2 \\ &= (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T(\mathbf{y} - \mathbf{m}) = \mathbf{G}(\mathbf{y} - \mathbf{m}),\end{aligned} \tag{14}$$

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 (7) and (14) that the genie bound is

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 = \mathbb{E}\|\mathbf{G}(\mathbf{J} - \mathbf{m})\|_2^2 \\ &= \mathbb{E}\left(\mathbf{tr}\left\{\mathbf{G}\mathbf{C}(\mathbf{J} - \mathbf{m})\mathbf{G}^T\right\}|\mathbf{A}_g\right),\end{aligned} \tag{15}$$

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

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 &= \mathbb{E}\left(\mathbf{tr}\left\{\sigma_J^2(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\right\}\right) \\ &= \sigma_J^2\mathbb{E}\left(\sum_{i=1}^{F}\lambda_i^{-1}(\mathbf{A}_g^T\mathbf{A}_g)\right),\end{aligned} \tag{16}$$

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 in Appendix, it can be verified that there exists a constant

$$c = \Theta(\sqrt{n/g(n)}) \tag{17}$$

[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.

such that each element in $c\mathbf{A}_g$ has finite mean and variance 1. We then re-write (16) as

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 &= \sigma_J^2 c^2 \mathbb{E}\left(\sum_{i=1}^{F} \lambda_i^{-1}\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)/L\right) && (18)\\ &\geq \sigma_J^2 c^2 \mathbb{E}\left(\frac{F^2/L}{\sum_{i=1}^{F} \lambda_i\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)}\right) && (19)\\ &\geq \sigma_J^2 c^2 \frac{F/L}{\mathbb{E}\left(\sum_{i=1}^{F} \lambda_i(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g))/F\right)}, && (20)\end{aligned}$$

where (19) follows from the Cauchy-Schwarz inequality, and (20) follows from the property of expectation.

According to Theorem of Universality for Bulk Convergence [19]–[21], 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

$$\mathbb{E}(\sum_{i=1}^{F} \lambda_i\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)/F) = \Theta\left(1\right). \quad (21)$$

Inserting (21) into (20) and using the fact that $L = \Theta(n)$ with high probability in Lemma A2 lead to

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 \geq \sigma_J^2 c^2 \Theta\left(F/n\right). \quad (22)$$

Inserting (17) into (22) yields

$$\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 \geq \Theta\left(\sigma_J^2 F/g(n)\right). \quad (23)$$

On the other hand, starting from (18), we have $\mathbb{E}\|\hat{\mathbf{x}}_g - \mathbf{x}_g\|_2^2 \leq \sigma_J^2 \mathbb{E}\left(\sum_{i=1}^{F} \lambda_{\min}^{-1}\left(\mathbf{A}_g^T\mathbf{A}_g\right)\right)$. Then, it follows from Lemma A4 that $\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 \leq \Theta\left(\sigma_J^2 F/g(n)\right)$, which is combined with (23) 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,

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 = \mathbb{E}\|\mathbf{G}\mathbf{J}\|_2^2 = \mathbb{E}\|\mathbf{G}(\mathbf{J}-\mathbf{m})+\mathbf{G}\mathbf{m}\|_2^2\\ &\geq \mathbb{E}\|\mathbf{G}(\mathbf{J}-\mathbf{m})\|_2^2 = \Theta\left(\sigma_J^2 F/g(n)\right), && (24)\end{aligned}$$

in which the last equality follows from (23). On the other hand, we have

$$\begin{aligned}&\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 = \mathbb{E}\|\mathbf{G}\mathbf{J}\|_2^2 \leq \mathbb{E}(\lambda_{\max}(\mathbf{G}^T\mathbf{G}))\mathbb{E}\|\mathbf{J}\|_2^2\\ &= \mathbb{E}(1/\lambda_{\min}(\mathbf{A}_g^T\mathbf{A}_g))\mathbb{E}\|\mathbf{J}\|_2^2 = \Theta\left((m_J^2+\sigma_J^2)n/g(n)\right), && (25)\end{aligned}$$

where the last inequality follows from Lemma A1, and the last equality follows from Lemmas A2 and A4. Combining (24) and (25) 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

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_g-\mathbf{x}_g\|_2^2 &= \mathbb{E}\|(\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T\mathbf{y}-\mathbf{x}_g\|_2^2\\ &= \mathbb{E}\|\mathbf{G}(\mathbf{B}_g-\mathbf{A}_g)\mathbf{x}_g\|_2^2, && (26)\end{aligned}$$

where $\mathbf{G} = (\mathbf{A}_g^T\mathbf{A}_g)^{-1}\mathbf{A}_g^T$. It follows from Lemma A1 in Appendix that $\lambda_{\min}(\mathbf{G}^T\mathbf{G})\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \|\mathbf{G}(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \lambda_{\max}(\mathbf{G}^T\mathbf{G})\|(\mathbf{B}_g - \mathbf{A}_g)\mathbf{x}_g\|_2^2$. 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

$$\frac{\mathbb{E}\|(\mathbf{B}_g-\mathbf{A}_g)\mathbf{x}_g\|_2^2}{g(n)+Fg(n)^2/n} \leq \mathbb{E}\|\mathbf{G}(\mathbf{B}_g–\mathbf{A}_g)\mathbf{x}_g\|_2^2 \leq \frac{\mathbb{E}\|(\mathbf{B}_g-\mathbf{A}_g)\mathbf{x}_g\|_2^2}{\Theta(g(n))}. \quad (27)$$

Next, we proceed to derive $\mathbb{E}\|(\mathbf{B}_g–\mathbf{A}_g)\mathbf{x}_g\|_2^2$ in (27). 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 A5, $s_{l,f}$ satisfies

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

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

$$\begin{aligned}&\mathbb{E}\|(\mathbf{B}_g-\mathbf{A}_g)\mathbf{x}_g\|_2^2 = \mathbb{E}(\sum_{l=1}^{L}(\sum_{f=1}^{F} s_{l,f}x_f)^2)\\ &= \Theta(n)\mathbb{E}((\sum_{f=1}^{F} s_{l,f}x_f)^2) \geq \Theta(n)(\mathbb{E}(\sum_{f=1}^{F} s_{l,f}x_f))^2\\ &= \Theta(n)F^2\Theta\left(h(n)^2/n^2\right)\mathbb{E}(x_f)^2\\ &= \Theta\left(F^2(m_L^2+\sigma_L^2)h(n)^2/n\right). && (28)\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}\|(\mathbf{B}_g-\mathbf{A}_g)\mathbf{x}_g\|_2^2 = \Theta(n)\mathbb{E}(\sum_{f=1}^{F} s_{l,f}x_f)^2\\ &\leq \Theta(n)\mathbb{E}(\sum_{f=1}^{F} s_{l,f}^2)\mathbb{E}(\sum_{f=1}^{F} x_f^2) = F^2\Theta(h(n))(m_L^2+\sigma_L^2). && (29)\end{aligned}$$

Combining (27), (28), and (29) 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

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

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 $\mathbb{E}\big(\mathbf{J}^T\mathbf{G}^T\mathbf{G}(\mathbf{B}_g-\mathbf{A}_g)\mathbf{x}\big)$ 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 see whether the inference errors predicted by the theoretical results match the numerical results. 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 $r=1$.

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 and we set that it yields a path length $h(n)=1.3g(n)$. Note that $g(n)$ and $h(n)$ should be carefully evaluated or measured for practical routing changing systems.

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.

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

We compute from (3) and (4) 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 (5) 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, 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; otherwise, it is increasing approximately on the order of $\sqrt{n}$. It is worth noting that routing changing induces higher genie bounds (i.e., more errors) for network inference, which also increase approximately on the order of $\sqrt{n}$ as observed in Fig. 4. We can also see from Fig. 4 that the combined strategy leads to the highest genie bound, yet still on the order of $\sqrt{n}$.

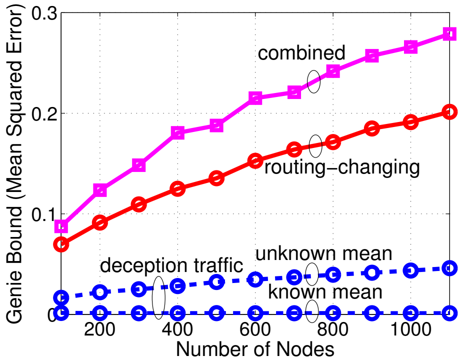


Fig. 4. Measured genie bounds under different strategies when $F=\lfloor\sqrt{n}\rfloor$.

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 $\Theta(n)\leq\mathcal{G}(\mathbf{x}_g|\mathcal{R})\leq\Theta(n^2)$.

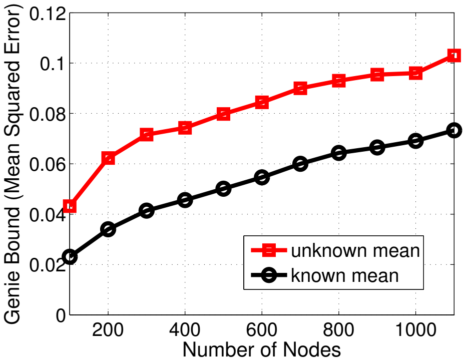


Fig. 5. Measured genie bounds under deception traffic when $F=n$.

Fig. 5 shows the measured genie bounds under the deception traffic strategy versus the number of nodes n. 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 n. We can see that all the genie bounds increase linearly, and the gaps between the three strategies are small.

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.

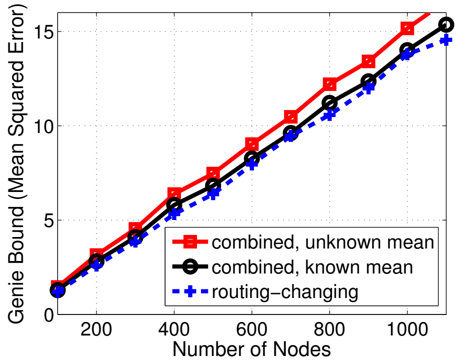


Fig. 6. Measured genie bounds under the routing changing and combined strategies in the $F = n$ scenario.

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 [18], 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.

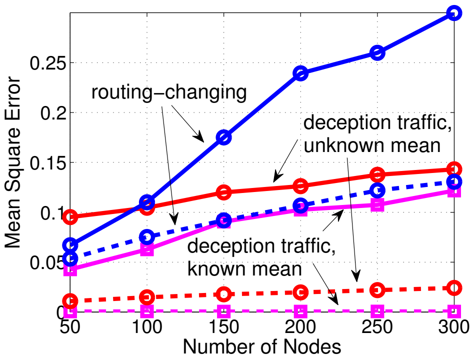


Fig. 7. Mean square error of the in-crowd algorithm under different strategies.

Fig. 7 shows the mean squared errors (in solid lines) of flow estimation based on the in-crowd algorithm under different proactive strategies. The genie bounds for these strategies are also drawn in dashed lines in Fig. 7. By comparing the solid and dash lines, we conclude that proactive strategies cause much more errors to practical algorithms (that are non-optimal and lead to performance penalties compared with the genie bound) for network inference. Thus, the theoretical analysis can serve as the at-least disruption benchmark for the impacts of proactive strategies on network inference in practice.

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

[11] 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.

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

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

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

[15] 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.

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

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

[18] 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.

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

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

[21] 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.

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

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

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

Appendix

In this appendix, we briefly prove lemmas used in analysis.

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 $\|\mathbf{X}\mathbf{a}\|_2^2 = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a} = \frac{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a}}{\mathbf{a}^T\mathbf{a}}\mathbf{a}^T\mathbf{a} = \frac{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a}}{\mathbf{a}^T\mathbf{a}}||\mathbf{a}||_2^2$, where $\frac{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a}}{\mathbf{a}^T\mathbf{a}}$ is called the Rayleigh quotient [22] 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., $\mathbb{P}(L = \Theta(n)) = 1 - \Theta(e^{-\Theta(1)n})$.

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. The total number of links in the network L can be written as $L = \sum_{i=1}^n l_i \sim$ Poisson$(n(\pi r^2\lambda - 1)/2)$. Thus, for some small positive constant $c_1 < \pi r^2\lambda - 1$, it follows from a Chernoff bound argument in [23] that

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

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

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

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 $\Theta\left(kn/h(n)\right)$. Thus, $\|\mathbf{J}\|_2^2$ at least has value

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

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 (i) the minimum eigenvalue $\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = \Theta(g(n))$ and (ii) the maximum eigenvalue $\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leq \Theta\left(g(n) + Fg(n)^2/n\right)$ asymptotically almost surely.

Proof: (i) It can be verified based on Lemma A5 that there exists a constant $c = \Theta(\sqrt{n/g(n)})$ 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 [19], the eigenvalue $\lambda_{\min}\left(L^{-1}(c\mathbf{A}_g)^T(c\mathbf{A}_g)\right)$ 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.

(ii) 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

$$\begin{aligned}\lambda_{\max}\left(\mathbf{X}^T\mathbf{X}\right) &\leq \lambda_{\max}\left(\mathbf{Y}^T\mathbf{Y}\right) + 2g(n)\lambda_{\max}\left(\mathbf{Y}^T\mathbf{Z}\right)/n\\ &\quad + g(n)^2\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right)/n^2. \qquad \text{(A1)}\end{aligned}$$

It follows from [19] that

$$\lambda_{\max}\left(\mathbf{Y}^T\mathbf{Y}\right) = \Theta(g(n)) \qquad \text{(A2)}$$

with high probability.

Then, we take a look at $\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right)$ in (A1). 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 [24] that, asymptotically almost surely, $\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right) = o((nF)^{1/p})$ for any $1 \leq p < 2$. Since $F \leq L = \Theta(n)$, we have

$$\frac{g(n)}{n}\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Y}\right) = o\left(\frac{F^{1/p}g(n)}{n^{1-\frac{1}{p}}}\right) \leq o\left(n^{\frac{2}{p}-1}g(n)\right). \qquad \text{(A3)}$$

Next, we consider $\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right)$ in (A1). Similar to $\left(\mathbf{Z}^T\mathbf{Y}\right)$, $\mathbf{Z}^T\mathbf{Z}$ is also of rank 1. Hence, we obtain

$$\lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right) = \mathbf{tr}\{\mathbf{Z}^T\mathbf{Z}\} = \sum_{l=1}^L\sum_{f=1}^F 1 = \Theta(Fn). \qquad \text{(A4)}$$

Inserting (A2), (A3), and (A4) into (A1) and letting $\xi = 2/p - 1$ yield $\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leq \Theta\left(n^\xi g(n) + Fg(n)^2/n\right)$ asymptotically almost surely for any arbitrarily small $\xi > 0$, which means that $\lambda_{\max}(\mathbf{X}^T\mathbf{X}) \leq \Theta\left(g(n) + Fg(n)^2/n\right)$. □

Lemma A5: Under Model 1, the probability that element $a_{i,j}$ in routing matrix $\mathbf{A}_g$ is 1 is $\mathbb{P}(a_{i,j} = 1) = \Theta(g(n)/n)$.

Proof: First, $\mathbb{P}(a_{i,j} = 1)$ 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, $\mathbb{P}(a_{i,j} = 1|l, f) = \binom{l-1}{f-1}/\binom{l}{f} = (l-1)!f!(l-f+1)!/(l!(f-1)!(l-f+1)!) = f/l$.

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

To proceed, write $\mathbb{E}(a_{i,j} = 1|f)$ as

$$\begin{aligned}\mathbb{E}(a_{i,j} = 1|f) &= \mathbb{E}(f/l|f) = f(1/\mathbb{E}(l|f) + O(\text{Var}(l|f)/\mathbb{E}(l|f)^3))\\ &= f/a(n) + O(1/n^2)\end{aligned}$$

where Var$(l|f)$ is the variance of l conditioned on f, and $a(n)$ is some function satisfying $a(n) = \Theta(n)$. The last equality holds due to the fact that Var$(l|f) = \Theta(n)$ as l follows the Poisson distribution as shown in Lemma A2.

Hence, we obtain

$$\mathbb{E}(\mathbb{E}(a_{i,j} = 1|f)) = \mathbb{E}(f/a(n)) = \Theta(g(n)/n), \qquad \text{(A5)}$$

which finishes the proof. □