On Covert Data Falsification Attacks on Distributed Detection Systems

Bhavya Kailkhura†, Yunghsiang S. Han†, Swastik Brahma∗, and Pramod K. Varshney∗
∗Department of EECS, Syracuse University, NY, 13244 USA
Email:{bkailkhu,skbrahma,varshney}@syr.edu
†Department of EE, National Taiwan University of Science and Technology, Taiwan, ROC
Email:yshan@mail.ntust.edu.tw

Abstract—In distributed detection systems, nodes make one bit decisions regarding the presence of a phenomenon and collaboratively make a global decision at the fusion center (FC). The performance of such systems strongly depends on the reliability of the nodes in the network. The robustness of distributed detection systems against attacks is of utmost importance for the functioning of distributed detection systems. The distributed nature of such systems makes them quite vulnerable to different types of attacks. In this paper, we introduce the problem of intelligent data falsification attacks on distributed detection systems. First, we propose a scheme to detect data falsification attacks and analytically characterize its performance. Next, we obtain the optimal attacking strategy from the point of view of a smart adversary to disguise itself from the proposed detection scheme while accomplishing its attack.

Index Terms—Distributed detection, Data falsification attack, Byzantines

I. INTRODUCTION

Distributed detection is a well studied topic in the detection theory literature [1]–[3]. Specifically, detection using the parallel topology has been considered extensively, wherein nodes make their local decisions regarding the underlying phenomenon and send it to the fusion center (FC), where a global decision is made. In recent years, security issues of such distributed networks are increasingly being studied. One typical attack on such networks is a data falsification attack [4]–[10]. We refer to a data falsification attacker as a Byzantine. Byzantines, in order to undermine the network performance, may alter their decisions with a certain probability prior to transmission to the FC. In other words, Byzantine attackers are those compromised nodes which send falsified decisions to degrade the system performance. In [6], [7], the authors considered the problem of determining the most effective Byzantine attacking strategy. The attacker’s only objective in these formulations is to degrade the detection performance as much as possible. We call such attacks, overt data falsification attacks. It was shown that when the fraction of Byzantines $\alpha \leq 0.5$, flipping their decision with probability $p = 1$ is the best strategy for Byzantines\(^1\). In other words, an attacker should attack with its full efficacy, $p = 1$, to degrade the network wide detection performance. Observe that, the overt data falsification attackers can be easily detected and discarded because attacking with full efficacy can very easily expose the adversary to the fusion center’s defense mechanism. However, knowing the existence of the defense mechanism, a smart adversary can manage to disguise its intention while accomplishing its attack.

In this paper, we introduce an intelligent data falsification attack on distributed detection systems, which aim for maximal damage by taking advantage of the knowledge of the defense mechanism. We assume that the adversary can make the maximum use of the knowledge of the defense mechanism to launch an attack. Thus, the results presented here provide the worst case assessment of the risk of data falsification attacks. The attack model in this paper has two parameters: the fraction of attackers ($\alpha$) in the network, and the flipping probability ($p$). These parameters determine the efficacy and the covertness of the attack. Here, the efficacy of an attack can be described by the degradation in the detection performance. The covertness of the attack can be described by the probability that the attacker will not be detected. The efficacy of attack increases with $p$ and $\alpha$. On the other hand, when the attack efficacy increases, the covertness decreases. Therefore, there is a trade-off between attack efficacy and covertness. Ideally, an attacker would prefer to maximize the attack efficacy while constraining its exposure to the defense mechanism, and, therefore, the problem becomes a constrained optimization problem. The main contributions of this paper are as follows.

1. We formulate covert data falsification attacks on distributed detection systems as a constrained optimization problem.
2. We propose a scheme to detect Byzantines and analytically evaluate its performance.
3. We obtain a closed-form expression for the optimal flipping probability that a smart adversary should use to minimize the detection performance while constraining its exposure to the defense mechanism.

The rest of the paper is organized as follows. Section II introduces our system model, including the problem formulation. In Section III, we propose and analyze a Byzantine detection scheme to be implemented at the FC. Section IV contains analysis of the covert data falsification attack problem along

\(^1\)Notice that, when $\alpha > 0.5$, all flipping probabilities $p$ which satisfy $\alpha p = 0.5$ will corresponds to the best strategy.
where b to the FC when its actual decision is q performance. The KLD between the distributions $P$ is evaluated to be:

$$D(r||q) = \sum_{j \in \{0,1\}} P(u_i = j|H_1) \log \frac{P(u_i = j|H_1)}{P(u_i = j|H_0)}.$$ 

The Byzantine nodes want to make the KLD as small as possible. The KLD under data falsification attack can be evaluated to be:

$$D(r||q) = (1 - \alpha p - P_d(1 - 2\alpha p)) \log \frac{1 - \alpha p - P_d(1 - 2\alpha p)}{1 - \alpha p - P_f(1 - 2\alpha p)} + (\alpha p + P_d(1 - 2\alpha p)) \log \frac{\alpha p + P_d(1 - 2\alpha p)}{\alpha p + P_f(1 - 2\alpha p)}.$$

### C. Problem Formulation

We formulate the problem of covert data falsification attacks as follows.

Knowing the existence of the defense mechanism, an intelligent adversary would minimize KLD while constraining its exposure to the defense mechanism. The covert data falsification attack problem can be formally stated as

$$\begin{align*}
\text{minimize} & \quad D(r||q) \\
\text{subject to} & \quad P^{iso}_B \leq \gamma \\
& \quad 0 \leq p \leq 1
\end{align*}$$

where $P^{iso}_B$ is the probability that a Byzantine is detected and isolated by the defense mechanism. An attacker tries to maintain $P^{iso}_B$ below a threshold $\gamma$ to constrain its exposure to the defense/detection scheme. Note that, the covert data falsification problem is dependent on the Byzantine detection scheme employed in the system. Hence, we next propose an efficient Byzantine detection scheme and solve the optimization problem based on the proposed scheme.

### III. An Efficient Byzantine Detection Scheme

In this section, we propose and analyze a Byzantine detection scheme to be implemented at the FC and obtain a closed form expressions to get a better insight into the problem.

#### A. Byzantine Detection Scheme

We assume that the FC observes the local decisions of each node over a time window $T$, which can be denoted by $U^i = [u^i_1, \ldots, u^i_T]$ for node $i$. We also assume that there is one honest anchor node with probabilities of detection $P^A_d$ and probabilities of false alarm $P^A_f$ present and known to the FC. We employ the anchor node to provide the gold standard which is used to detect whether or not other nodes are Byzantines. The FC can also serve as an anchor node in that it can directly observe the phenomenon and make a decision. We denote the Hamming distance between reports of the anchor node and node $i$ over the time window $T$ by $d^A_i = ||U^A - U^i||$, that is the number of elements that are different between $U^A$ and $U^i$. Since the FC is aware of the fact that Byzantines might be present in the network, it compares $d^A_i$ to a threshold $\eta$ to make the decision regarding the presence of the Byzantines. In this paper, we counter the data falsification attack by isolating or cutting-off those nodes from the information fusion process whose distance $d^A_i$ is greater than a fixed threshold $\eta$. The probability that a Byzantine is isolated at the end of the time window $T$, $P^{iso}_B$, is a function of the parameter $\eta$, which is under the control of the FC and parameters $(p, \alpha)$, which are under the control of the attacker.

#### B. Performance Analysis

As aforementioned, local decisions of the nodes are compared to the decisions of the anchor node over a time window.
of length $T$. The probability that an honest node makes a decision that is different from the anchor node is given by
\[ P_{AH}^{iso} = P(D_i^A > \eta) = \sum_{j=\eta+1}^{T} \binom{T}{j} (P_{d_{diff}}^{AB})^j (1 - P_{d_{diff}}^{AB})^{T-j}. \] (2)

When $T$ is large enough, by using the Normal approximation, we have
\[ P_{iso} = Q \left( \frac{\eta - TP_{d_{diff}}^{AB}}{\sqrt{(TP_{d_{diff}}^{AB})(1 - P_{d_{diff}}^{AB})}} \right). \] (3)

Similarly, the isolation probability for an honest node can be obtained as
\[ P_{iso}^{H} = Q \left( \frac{\eta - TP_{d_{diff}}^{AB}}{\sqrt{(TP_{d_{diff}}^{AB})(1 - P_{d_{diff}}^{AB})}} \right). \] (4)

The optimal threshold, $\eta^*$, is selected by constraining the isolation probability of honest nodes such that $P_{iso}^{H} \leq \kappa$. That is,
\[ \eta^* = Q^{-1}(\kappa) \sqrt{(TP_{d_{diff}}^{AB})(1 - P_{d_{diff}}^{AB})} + TP_{d_{diff}}^{AB}. \] (5)

In Figure 1, we plot the achievable $P_{iso}^{B}$ values as a function of the flipping probability $p$, for $P_0 = P_1 = 0.5$, $(P_d, P_f) = (0.8, 0.2)$, $T = 20$, and $\kappa = 0.3$. It can be observed from this figure that $P_{iso}^{B}$ increases with flipping probability $p$.

IV. OPTIMAL COVERT DATA FALSIFICATION ATTACK

In this section, we first discuss a property of KLD with respect to the flipping probability $p$ that helps us formulate an equivalent and more tractable optimization problem. Next, the solution of the covert data falsification attack is provided based on the equivalent problem formulation.

A. Problem Analysis

We discuss a property of KLD with respect to the flipping probability $p$. Figure 2 illustrates the result that $KLD$ is a monotonically increasing function of $t = (1 - 2\alpha p)^2$ when $P_d = 0.6$ and $P_f = 0.2$. We next prove that it is true for any $P_d > 0.5$ and $P_f < 0.5$.

**Lemma 1:** KLD under a data falsification attack is a monotonically increasing function of $(1 - 2\alpha p)^2$.

**Proof:** Let $t = (1 - 2\alpha p)^2$ or $\alpha p = \frac{1 - \sqrt{t}}{2}$. When $\alpha p = \frac{1 - \sqrt{t}}{2}$, we have
\[ D(r|q) = \frac{1 + \sqrt{t}}{2} P_d \sqrt{t} \log \frac{1 + \sqrt{t} - 2P_f \sqrt{t}}{1 + \sqrt{t} - 2P_f \sqrt{t}} + \frac{1 - \sqrt{t}}{2} + P_d \sqrt{t} \log \frac{1 - \sqrt{t} + 2P_d \sqrt{t}}{1 - \sqrt{t} + 2P_f \sqrt{t}}. \]
\[
\frac{dD}{dt} = \frac{2P_d - 1}{4\sqrt{t}} \left[ \left( 1 - \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} \right) - \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} \right] = \frac{2P_d - 1}{4\sqrt{t}} \left[ \log \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} - \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} \right] \]
\[
= \frac{2P_d - 1}{4\sqrt{t}} \left[ \log \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} - \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} \right] + \frac{2P_d - 1}{4\sqrt{t}} \left[ \log \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} - \frac{1 + \sqrt{t}(2P_d - 1)}{1 + \sqrt{t}(2P_d - 1)} \right].
\]
\[
(6)
\]

where \( P^* = \max(0, M(1)) \). (7)

\[
M(i) = \frac{T(2\eta^* + (Q^{-1}(\gamma))^2) + (1 - \eta^*)T(2\eta^* + (Q^{-1}(\gamma))^2) + 4(\eta^*)^2T(2\eta^* + (Q^{-1}(\gamma))^2)}{2T(2\eta^* + (Q^{-1}(\gamma))^2T)}
\]

for \( i = 0, 1 \)

This inequality is true from (8) by using the fact that if \( x > y \) then \( \log(x) > \log(y) \). Inequalities (8) and (9) imply that \( \frac{dD}{dt} > 0 \), which completes the proof. ■

Based on Lemma 1, the covert data falsification optimization problem given in (1) is equivalent to

\[
\text{minimize} \quad (1 - 2\alpha p)^2
\]

subject to \( P_B^{iso} \leq \gamma \)

\[
0 \leq p \leq 1.
\]

\( (10) \)

B. Optimal Attacking Strategy

Solution of the optimization problem (10) based on the proposed Byzantine-detection method is obtained analytically and is given as follows:

1) Case 1: When \( \alpha \leq 0.5 \): If \( P_B^{iso}(\alpha, p = 1) \leq \gamma \), covertness constraint is not active and the attacker should flip with probability \( p = 1 \), similar to the no constraint case. When \( P_B^{iso}(\alpha, p = 1) > \gamma \), covertness constraint is active. In this case, the attacker will start decreasing the flipping probability until \( P_B^{iso}(\alpha, p = 1) = \gamma \). For a given \( \gamma \) and \( \eta^* \) as given in (5), the optimal flipping probability \( p^* \) is obtained by setting \( P_B^{iso} = \gamma \) as given in (7).

Case 2: When \( \alpha > 0.5 \): If \( P_B^{iso}(\alpha, p = 1) \leq \gamma \), covertness constraint is not active and the attacker would flip with probability \( p = \frac{1}{2\alpha} \), similar to the no constraint case. Otherwise, it would flip with probability \( p^* \) given in (7).

Next, we analyze the minimum fraction of Byzantines needed to make \( KLD = 0 \) or to blind the FC. By Lemma 1, \( KLD = 0 \) only when \( 1 - 2\alpha p^2 = 0 \).

2) Case 1: When \( \alpha \leq 0.5 \): If \( P_B^{iso}(\alpha, p = 1) \leq \gamma \), the minimum fraction of Byzantines needed to make \( KLD = 0 \) is 50\% or \( \alpha_{blind} = 0.5 \); otherwise, the attacker cannot blind the FC.

Case 2: When \( \alpha > 0.5 \): If \( P_B^{iso}(\alpha, p = 1) \leq \gamma \), the attacker can blind the FC; otherwise it cannot. In this case, minimum fraction of Byzantines needed to make \( KLD = 0 \) is \( \alpha_{blind} = \frac{1}{2p^*} \), where \( p^* \) as given in (7).
of Byzantines in the network. We plot removed from the fusion process as a function of the fraction of Byzantines, \( \gamma \) covertness constraints, \( \alpha \) function of the fraction of Byzantine attackers \( p \) is present in the network. Prior probability of the hypotheses are assumed to be equal, \( P_0 = P_1 = 0.5 \). The FC observes local decision of the nodes over a \( T = 15 \) time window. Reputation metric \( \eta \) has been chosen such that the probability of an honest node being removed from the process at the end of the time window \( T \) is \( P_{iso}^D < 0.3 \), i.e., \( \kappa = 0.3 \).

In Figure 3, we plot the optimal flipping probability, \( p^* \), as a function of fraction of the Byzantines, \( \alpha \), for different covertness constraints, \( \gamma \). Observe that, when there is no constraint on \( \gamma \) \([6], [7]\), flipping with probability \( p = 1 \) is the best strategy for \( \alpha \leq 0.5 \) and for \( \alpha > 0.5 \), all flipping probabilities \( p \) which satisfy \( \alpha p = 0.5 \) are the best strategies. For \( \alpha \leq 0.5 \), all three covertness constraints considered, \( \gamma = [0.55, 0.65, 0.75] \), are less than \( P_{iso}^D(\alpha, p = 1) = 0.8727 \), (see Figure 4), thus covertness constraints are active. Therefore, the attacker chooses flipping probability \( p^* \) given in (7) which is less than 1. For \( \alpha > 0.5 \), the attacker continues to flip with probability \( p^* \) until \( P_{iso}^D(\alpha, p = \frac{1}{2\alpha}) > \gamma \) and after that point, the attacker flips with probability \( \frac{1}{2\alpha} \), similar to the no constraint case. Notice that, for \( \gamma = 0.75 \), \( P_{iso}^D(\alpha, p = \frac{1}{2\alpha}) < \gamma \), \( \forall \alpha \), therefore the optimal flipping probability \( p^* \) is the same for all \( 0 \leq \alpha \leq 1 \).

Figure 4 shows the probability of a Byzantine node being removed from the fusion process as a function of the fraction of Byzantines in the network. We plot \( P_{iso}^D \) for different covertness constraints, \( \gamma \). (Byzantines flip their decision with probability \( p^* \) as shown in Figure 3). We observe that the isolation probability of a Byzantine, \( P_{iso}^D \), is a non-increasing function of the fraction of Byzantine attackers \( \alpha \) and follows a similar pattern as in Figure 3.

Next, in Figure 5, we analyze the effect of fraction of the Byzantines on detection performance. We plot minimum K-L divergence, \( \min D \), as a function of fraction of the Byzantines \( p \) for different covertness constraints, \( \gamma \). Observe that, in the absence of the covertness constraint, 50% Byzantine nodes are needed to make \( D = 0 \) or blind the fusion center; however, \( P_{iso}^D \) is very high for this case and Byzantine nodes can be easily caught by the FC. For covertness constraints, \( \gamma = [0.55, 0.65] \), the attacker can still make \( D = 0 \) or blind the fusion center; however, the fraction of Byzantines needed to blind the FC increases. When covertness constraint \( \gamma = 0.75 \), the attacker cannot blind the FC; however, it can make KLD very low, \( D < 0.025 \), and still maintains high covertness.

Numerical results presented in this section suggest that there is a trade-off between attack efficacy and covertness. By attacking with flipping probability \( p^* \), an adversary can degrade the detection performance while constraining its exposure to the defense mechanism.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced an intelligent data falsification attack on distributed detection systems. First, we proposed a data falsification detection scheme and analytically characterized its performance. Next, we obtained a closed form
expression for the optimal flipping probability \( p^* \) that a smart adversary could use to disguise its intention while accomplishing its attack. By attacking with flipping probability \( p^* \), an adversary maximizes the attack efficacy while constraining its exposure to the defense mechanism. There are still many interesting questions that remain to be explored in the future work such as an analysis of the scenario where both the FC and the Byzantine attacker act in a strategic manner to optimize their own utilities. The optimal strategies which maximize their performance in finite cycles of data fusion can also be investigated.

ACKNOWLEDGEMENT

This work was supported in part by ARO under Grant W911NF-09-1-0244, AFOSR under Grants FA9550-10-1-0458, FA9550-10-1-0263 and National Science Council of Taiwan, under grants NSC 99-2221-E-011-158-MY3, NSC 101-2221-E-011-069-MY3. Han’s work was completed during his visit to Syracuse University from 2012 to 2013.

REFERENCES