

Detection of Malicious Agents in Social Learning

Valentina Shumovskaia, Mert Kayaalp, and Ali H. Sayed

École Polytechnique Fédérale de Lausanne (EPFL)

Abstract—Social learning is a non-Bayesian framework for distributed hypothesis testing aimed at learning the true state of the environment. Traditionally, the agents are assumed to receive observations conditioned on the same true state, although it is also possible to examine the case of heterogeneous models across the graph. One important special case is when heterogeneity is caused by the presence of malicious agents whose goal is to move the agents towards a wrong hypothesis. In this work, we propose an algorithm that allows to discover the true state of every individual agent based on the *sequence* of their beliefs. In so doing, the methodology is also able to locate malicious behavior.

Index Terms—Social learning, hypotheses testing, inverse modeling, diffusion strategy, adaptive learning, anomaly detection, malicious agent.

I. INTRODUCTION AND RELATED WORK

Social learning algorithms [1]–[12] solve the distributed hypothesis problem in a non-Bayesian fashion. These algorithms learn the underlying true state of nature by observing streaming data arriving at the agents and conditioned on that state. The key difference with Bayesian solutions [13]–[15] is that social learning does not require each node to know the full graph topology or likelihood models used by every other node (including non-neighboring nodes). These features enable fully decentralized implementations. Some applications of the social learning framework include detection problems by sensor networks [16], [17] or distributed machine learning [18], [19]. The framework can also be used to describe how users form their opinions over social graphs [20].

Under social learning, and in order to learn the truth, agents update their beliefs (or confidences) on each possible hypothesis, ensuring that the total confidence adds up to 1. In general, at every time instant, each agent receives an observation conditioned on the state of the environment and uses the local likelihood model to perform a local Bayesian update starting from its current belief vector. This step is followed by a communication stage, during which agents exchange beliefs with their neighbors and fuse the received information with their own opinions. These two steps of local Bayesian updates and local consultation are repeated until convergence is attained.

Many existing works on social learning assume that the observations received by each agent arise from *one* true state of the environment. Some other works study the case of nonhomogeneous models across the agents. For example, the works [21], [22] focus on community structure networks where each community has its own truth. The main conclusion is

that social learning methods are quite robust in the presence of malicious agents (which are defined as agents that feed observations that arise from a different state than the rest of the network). Some works suggest defense mechanisms against such agents [23], [24]. The robustness means that the malicious agents are forced to converge to the same conclusion as the rest of the network. This fact makes it impossible to identify malicious (or dysfunctional) agents, based solely on their observed belief.

In this work, we develop a centralised algorithm for identifying the true state associated with each agent, even when the final belief of an agent may be pointing toward another conclusion due to the interactions over the graph. In this way, the method is able to identify malicious agents as well. There is no question that this is an important question that deserves attention [25]–[38]. For instance, over social networks, it is critical to identify users that have unwarranted intentions and aim to force the network to reach erroneous conclusions [31]–[33], as well as to discover trolls [34]–[36] and measure their impact on performance [25]. The same techniques can be used to locate malfunctioning agents, e.g., [27].

II. SOCIAL LEARNING MODEL

A set of agents \mathcal{N} builds confidences on each hypothesis θ from a finite set Θ through interactions with the environment and among the agents. The agents communicate according to a fixed combination matrix $A \in [0, 1]^{\mathcal{N} \times \mathcal{N}}$, where each nonzero element $a_{\ell,k} > 0$ indicates a directed edge from agent ℓ to agent k and defines the level of trust that agent k gives to information arriving from agent ℓ . Each agent k assigns a total confidence level of 1 to its neighbors. This assumption makes the combination matrix A left stochastic, i.e.,

$$\sum_{\ell \in \mathcal{N}} a_{\ell,k} = 1, \quad \forall k \in \mathcal{N} \quad (1)$$

Another common assumption, ensuring global truth learning for homogeneous environments, is that A is strongly connected. This implies the existence of at least one self-loop with a positive weight and a path with positive weights between any two nodes [39]. This condition allows us to apply the Perron-Frobenius theorem [40, Chapter 8], [41], which ensures that the power matrix A^s converges exponentially to $u\mathbf{1}^T$ as $s \rightarrow \infty$. Here, u is the Perron eigenvector of A associated with the eigenvalue at 1 and is normalized as follows:

$$Au = u, \quad u_{\ell} > 0, \quad \sum_{\ell \in \mathcal{N}} u_{\ell} = 1. \quad (2)$$

Each agent assigns an initial *private* belief $\mu_{k,0}(\theta) \in [0, 1]$ to each hypothesis $\theta \in \Theta$, forming a probability mass function

with the total confidence summing up to 1, i.e., $\sum_{\theta} \mu_{k,0}(\theta) = 1$. To avoid excluding any hypothesis from the beginning, we assume that each component of the belief vector $\mu_{k,0}$ is positive. Subsequently, agents iteratively update their belief vectors by interacting both with the environment and with their neighbors. At each time instance i , agent k receives an observation from the environment conditioned on its true state, denoted by $\zeta_{k,i} \sim L_k(\zeta|\theta_k^*)$. In this notation, the observation $\zeta_{k,i}$ arises from the likelihood model $L_k(\zeta|\theta_k^*)$, which is parameterized by the unknown model θ_k^* . For example, the entire network may be following the same and unique model θ^* , while a few malicious agents may be following some other model $\theta \neq \theta^*$. The observations $\{\zeta_{k,i}\}$ are assumed to be independent and identically distributed (i.i.d.) over time. The local Bayesian update performed by agent k at time i takes the following form [7]:

$$\psi_{k,i}(\theta) = \frac{L_k^\delta(\zeta_{k,i} | \theta) \mu_{k,i-1}^{1-\delta}(\theta)}{\sum_{\theta' \in \Theta} L_k^\delta(\zeta_{k,i} | \theta') \mu_{k,i-1}^{1-\delta}(\theta')}, \quad \forall k \in \mathcal{N}, \quad (3)$$

where $\delta \in (0, 1)$ plays the role of an adaptation parameter and it controls the importance of the newly received observation relative to the information learned from past interactions. The denominator in (3) serves as a normalization factor, ensuring that the resulting $\psi_{k,i}$ is a probability mass function. We refer to $\psi_{k,i}$ as the *public* (or intermediate) belief due to the next communication step, which involves a geometric averaging computation [2], [4], [9]:

$$\mu_{k,i}(\theta) = \frac{\prod_{\ell \in \mathcal{N}_k} \psi_{\ell,i}^{a_{\ell k}}(\theta)}{\sum_{\theta' \in \Theta} \prod_{\ell \in \mathcal{N}_k} \psi_{\ell,i}^{a_{\ell k}}(\theta')}, \quad \forall k \in \mathcal{N}. \quad (4)$$

At each iteration i , each agent k estimates its true state θ_k^* based on the belief vector (either private or public) by selecting the hypothesis with the highest confidence:

$$\hat{\theta}_{k,i} \triangleq \arg \max_{\theta \in \Theta} \mu_{k,i}(\theta). \quad (5)$$

In the homogeneous environment case [2], [4], [7], [9], i.e., when $\theta_k^* = \theta^*$ for each k , it can be proved that every agent finds the truth asymptotically with probability 1. The work [22] considers non-homogeneous environments with community-structured graphs; it establishes that, as $\delta \rightarrow 0$, the entire network converges to *one* solution that best describes the data, while in contrast, a larger δ activates the mechanism of *local* adaptivity. Thus, with a larger δ , each individual agent focuses more on its immediate neighborhood than on the entire network.

III. INVERSE MODELING

In this section, we explain how we can identify malicious agents (or the true state θ_k^* for each agent) by observing sequences of public beliefs. Importantly, we will not assume knowledge of the combination matrix A .

To begin with, we introduce the following common assumption, essentially requiring the observations to share the same support region [8], [20], [42].

Assumption 1 (Bounded likelihoods). *There exists a finite constant $b > 0$ such that for all $k \in \mathcal{N}$:*

$$\left| \log \frac{L_k(\zeta | \theta)}{L_k(\zeta | \theta')} \right| \leq b \quad (6)$$

for all $\theta, \theta' \in \Theta$ and ζ . ■

Now, consider a *sequence* of public beliefs measured closer to the steady state:

$$\{\psi_{k,i}\}_{i \gg 1}, \quad k \in \mathcal{N} \quad (7)$$

When an agent cannot distinguish between θ_k^* and another θ due to $L_k(\theta_k^*) = L_k(\theta)$, we will treat this θ as a valid model for the agent as well. To accommodate this possibility, we define Θ_k^* as the optimal hypotheses subset for each individual agent, denoted by $\Theta_k^* = \{\theta_k^*\} \cup \{\theta \neq \theta_k^* \mid L_k(\theta) = L_k(\theta_k^*)\}$. Then, we reformulate the problem by stating that our aim is to recover the optimal hypotheses subset for each agent:

$$\{\Theta_k^*\}, \quad k \in \mathcal{N}. \quad (8)$$

We denote the level of informativeness of any pair of hypotheses $\theta, \theta' \in \Theta$ at each agent k by:

$$d_k(\theta, \theta') \triangleq \mathbb{E}_{\zeta_k \sim L_k(\theta_k^*)} \log \frac{L_k(\zeta_k | \theta)}{L_k(\zeta_k | \theta')} \quad (9)$$

It is clear that this value is equal to zero if both θ and θ' belong to the optimal subset Θ_k^* . Additionally, $d_k(\theta_k^*, \theta)$ will be positive for any $\theta \notin \Theta_k^*$, since

$$d_k(\theta_k^*, \theta) = D_{\text{KL}}(L_k(\theta_k^*) \parallel L_k(\theta)) > 0 \quad (10)$$

and, in turn, $d_k(\theta, \theta_k^*)$ is always negative:

$$d_k(\theta, \theta_k^*) = -D_{\text{KL}}(L_k(\theta) \parallel L_k(\theta_k^*)) < 0 \quad (11)$$

Here, D_{KL} denotes the Kullback-Leibler divergence between two distributions:

$$D_{\text{KL}}(L_k(\theta_k^*) \parallel L_k(\theta)) \triangleq \mathbb{E}_{\zeta \sim L_k(\zeta|\theta_k^*)} \log \frac{L_k(\zeta | \theta_k^*)}{L_k(\zeta | \theta)} \quad (12)$$

Properties (10)–(11) allow us to conclude that the optimal hypotheses subset Θ_k^* consists of all θ for which:

$$\Theta_k^* = \{\theta: d_k(\theta, \theta') \geq 0, \quad \forall \theta' \in \Theta\} \quad (13)$$

Our aim is to develop an algorithm that learns Θ_k^* based on the available information (7).

In [42, Appendix A], it was shown that the adaptive social learning iterations (3)–(4) can be expressed in the following compact linear form:

$$\mathbf{\Lambda}_i = (1 - \delta) A^\top \mathbf{\Lambda}_{i-1} + \delta \mathcal{L}_i \quad (14)$$

where $\mathbf{\Lambda}_i$ and \mathcal{L}_i are matrices of size $|\mathcal{N}| \times (|\Theta| - 1)$, and for each k and j , their entries take the log-ratio form:

$$[\mathbf{\Lambda}_i]_{k,j} \triangleq \log \frac{\psi_{k,i}(\theta_0)}{\psi_{k,i}(\theta_j)}, \quad [\mathcal{L}_i]_{k,j} \triangleq \log \frac{L_k(\zeta_{k,i} | \theta_0)}{L_k(\zeta_{k,i} | \theta_j)}. \quad (15)$$

for any ordering $\Theta = \{\theta_0, \dots, \theta_{|\Theta|-1}\}$. The expectation of

\mathcal{L}_i , relative to the observations $\{\zeta_{k,i}\}_k$, is given by:

$$[\bar{\mathcal{L}}]_{k,j} \triangleq [\mathbb{E}\mathcal{L}_i]_{k,j} = D_{\text{KL}}(L_k(\theta_k^*) \parallel L_k(\theta_j)) - D_{\text{KL}}(L_k(\theta_k^*) \parallel L_k(\theta_0)), \quad (16)$$

and it allows us to rewrite (9) in a slightly different manner:

$$d_k(\theta_{j_1}, \theta_{j_2}) = [\bar{\mathcal{L}}]_{k,j_2} - [\bar{\mathcal{L}}]_{k,j_1} \quad (17)$$

Furthermore, it was shown in [20] that we can estimate $\bar{\mathcal{L}}$ by utilizing the publicly exchanged beliefs with the following accuracy [20, Theorem 2]:

$$\begin{aligned} & \limsup_{i \rightarrow \infty} \mathbb{E} \|\hat{\mathcal{L}}_i - \bar{\mathcal{L}}\|_{\text{F}}^2 \\ & \leq \frac{1}{M} \text{Tr}(\mathcal{R}_{\mathcal{L}}) + O(\mu/\delta^2) + O(1/\delta^5 M^2) \end{aligned} \quad (18)$$

where μ is a small positive learning rate for a stochastic gradient implementation, M is a batch size of data used to compute the estimate $\hat{\mathcal{L}}_i$, and $R_{\mathcal{L}} \triangleq \mathbb{E}(\mathcal{L}_i - \bar{\mathcal{L}})(\mathcal{L}_i - \bar{\mathcal{L}})^{\top}$. Thus, the informativeness (17) can be estimated by using

$$\hat{d}_k(\theta_{j_1}, \theta_{j_2}) = [\hat{\mathcal{L}}]_{k,j_2} - [\hat{\mathcal{L}}]_{k,j_1} \quad (19)$$

where $\hat{\mathcal{L}}$ is the estimate of $\bar{\mathcal{L}}$ from the last available iteration. Based on (13), we can now identify the optimal hypotheses subset Θ_k^* defined in (13) as follows:

$$\hat{\Theta}_k \triangleq \arg \max_{\theta_{j_1}, \theta_{j_2}} \mathbb{I} \left\{ \hat{d}_k(\theta_{j_1}, \theta_{j_2}) > 0 \right\} \quad (20)$$

where $\mathbb{I}\{x\}$ is an indicator function that assumes the value 1 when its argument is true and is 0 otherwise.

We list the procedure in Algorithm 1, including the part related to estimating (18) by using [20, Algorithm 1].

The following result establishes the probability of error.

Theorem 1 (Probability of error). *The probability of choosing a wrong hypothesis $\theta \notin \Theta_k^*$ for agent $k \in \mathcal{N}$ is upper bounded by:*

$$\begin{aligned} \mathbb{P} \left\{ \theta \in \hat{\Theta}_k \right\} & \leq \frac{2}{M} \text{Tr}(\mathcal{R}_{\mathcal{L}}) \sum_{\theta^* \in \Theta_k^*} D_{\text{KL}}^{-1}(L_k(\theta^*) \parallel L_k(\theta)) \\ & + O(\mu/\delta^2) + O(1/\delta^5 M^2) \end{aligned} \quad (21)$$

Proof. First, we upper bound the probability using the definition of $d(\cdot, \cdot)$ and its estimate from (9) and (19), along with the properties of probability. For any $\theta_j \notin \Theta_k^*$, we have that:

$$\begin{aligned} \mathbb{P} \left\{ \theta_j \in \hat{\Theta}_k \right\} & \leq \mathbb{P} \left\{ \exists \theta_k^* \in \Theta_k^* : \hat{d}_k(\theta_k^*, \theta_j) < 0 \right\} \\ & \leq \sum_{\theta_k^* \in \Theta_k^*} \mathbb{P} \left\{ \hat{d}_k(\theta_k^*, \theta_j) < 0 \right\} \end{aligned} \quad (22)$$

Next, we estimate the probability of $\hat{d}_k(\theta_k^*, \theta_j)$ being negative for some fixed θ_j and θ_k^* using (19), while denoting j_k^* as the index of θ_k^* :

$$\begin{aligned} \mathbb{P} \left\{ \hat{d}_k(\theta_k^*, \theta_j) < 0 \right\} & = \mathbb{P} \left\{ [\hat{\mathcal{L}}]_{k,j} - [\hat{\mathcal{L}}]_{k,j_k^*} < 0 \right\} \\ & = \mathbb{P} \left\{ [\hat{\mathcal{L}}]_{k,j_k^*} - [\bar{\mathcal{L}}]_{k,j_k^*} - \left([\hat{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j} \right) \right. \\ & \quad \left. > [\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*} \right\} \end{aligned}$$

Algorithm 1: Inverse learning of heterogeneous states

Data: At each time i : $\{\psi_{k,i}(\theta)\}_{k \in \mathcal{N}}, \delta$

Result: Estimated combination matrix \mathbf{A} ;

Estimated expected log-likelihood ratios $\hat{\mathcal{L}}$;

Estimated true state of each agent, $\hat{\Theta}_k$.

initialize $\mathbf{A}_0, \hat{\mathcal{L}}_0$

repeat

 Compute matrices Λ_i :

for $k \in \mathcal{N}, j = 1, \dots, |\Theta|$ **do**

$$[\Lambda_i]_{k,j} = \log \frac{\psi_{k,i}(\theta_0)}{\psi_{k,i}(\theta_j)}$$

 Combination matrix update [20]:

$$\begin{aligned} \mathbf{A}_i & = \mathbf{A}_{i-1} + \mu(1-\delta) \left(\Lambda_{i-1} - \frac{1}{M} \sum_{j=i-M}^{i-1} \Lambda_{j-1} \right) \\ & \quad \times \left(\Lambda_i^{\top} - (1-\delta) \Lambda_{i-1}^{\top} \mathbf{A}_{i-1} - \delta \hat{\mathcal{L}}_{i-1}^{\top} \right). \end{aligned}$$

 Log-likelihoods matrix update:

$$\hat{\mathcal{L}}_i = \frac{1}{\delta M} \sum_{j=i-M+1}^i \left(\Lambda_j - (1-\delta) \mathbf{A}_i^{\top} \Lambda_{j-1} \right)$$

$i = i + 1$

until sufficient convergence;

 Informativeness estimate for all agents $k \in \mathcal{N}$ and pairs of hypotheses $\theta_{j_1}, \theta_{j_2} \in \Theta$:

$$\hat{d}_k(\theta_{j_1}, \theta_{j_2}) = [\hat{\mathcal{L}}_i]_{k,j_2} - [\hat{\mathcal{L}}_i]_{k,j_1}$$

 Optimal hypotheses set estimate for all agents $k \in \mathcal{N}$:

$$\hat{\Theta}_k \triangleq \arg \max_{\theta_{j_1}, \theta_{j_2}} \mathbb{I} \left\{ \hat{d}_k(\theta_{j_1}, \theta_{j_2}) > 0 \right\}$$

$$\begin{aligned} & = 1 - \mathbb{P} \left\{ [\hat{\mathcal{L}}]_{k,j_k^*} - [\bar{\mathcal{L}}]_{k,j_k^*} - \left([\hat{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j} \right) \right. \\ & \quad \left. \leq [\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*} \right\} \\ & \leq 1 - \mathbb{P} \left\{ \left| [\hat{\mathcal{L}}]_{k,j_k^*} - [\bar{\mathcal{L}}]_{k,j_k^*} \right| + \left| [\hat{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j} \right| \right. \\ & \quad \left. \leq [\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*} \right\} \\ & \leq 1 - \mathbb{P} \left\{ \left| [\hat{\mathcal{L}}]_{k,j_k^*} - [\bar{\mathcal{L}}]_{k,j_k^*} \right| \leq [\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*} \right\} \\ & \quad \times \mathbb{P} \left\{ \left| [\hat{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j} \right| \leq [\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*} \right\} \end{aligned} \quad (23)$$

We can transform the result from [42, Theorem 2] (18) into:

$$\mathbb{E} \left| [\hat{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j} \right| \leq \frac{1}{M} \text{Tr}(\mathcal{R}_{\mathcal{L}}) + O(\mu/\delta^2) + O(1/\delta^5 M^2) \quad (24)$$

By Markov's inequality [41], for any $a > 0$:

$$\begin{aligned} & \mathbb{P} \left(\left| [\hat{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j} \right| \leq a \right) \\ & \geq 1 - \frac{1}{aM} \text{Tr}(\mathcal{R}_{\mathcal{L}}) + O(\mu/\delta^2) + O(1/\delta^5 M^2) \end{aligned} \quad (25)$$



Fig. 1: Example of images from the MIRO dataset for classes “bus” and “car”.

Also, by the definition of KL divergence we have that:

$$[\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*} = D_{\text{KL}}(L_k(\theta_k^*) || L_k(\theta_j)) > 0. \quad (26)$$

Thus, (23) can be upper bounded by:

$$\begin{aligned} & \mathbb{P}\left\{\widehat{d}_k(\theta_k^*, \theta_j) < 0\right\} \\ & \leq 1 - \left(1 - \frac{\frac{1}{M}\text{Tr}(\mathcal{R}_{\mathcal{L}}) + O(\mu/\delta^2) + O(1/\delta^5 M^2)}{[\bar{\mathcal{L}}]_{k,j} - [\bar{\mathcal{L}}]_{k,j_k^*}}\right)^2 \\ & \approx 2M^{-1}\text{Tr}(\mathcal{R}_{\mathcal{L}}) D_{\text{KL}}^{-1}(L_k(\theta_k^*) || L_k(\theta)) \\ & \quad + O(\mu/\delta^2) + O(1/\delta^5 M^2) \end{aligned} \quad (27)$$

using the Taylor’s expansion for any small x , namely, $(1+x)^2 = 2x + O(x^2)$.

Combining (22) with (27) we get the desired statement. ■

IV. COMPUTER EXPERIMENTS

In this section, we consider the image dataset MIRO (Multi-view Images of Rotated Objects) [43], which contains objects of different classes from different points of view – see Fig. 1. For each class, there are 10 objects, and each of the objects has 160 different perspectives.

A network of agents wishes to solve a binary hypotheses problem to distinguish between states θ_0 corresponding to the class “bus” and θ_1 corresponding to the class “car”. Each agent has its own convolutional neural network (CNN) classifier. These CNNs are trained to distinguish classes θ_0 and θ_1 by observing only a part of the image, similar to the approach in [18], [19]. Each image measures 224×224 pixels, and each agent observes a section of size 112×112 pixels, situated in different regions of the image. We illustrate the observation map in Fig. 2a. The CNN architecture consists of three convolutional layers: 6 output channels, 3×3 kernel, followed by ReLU and 2×2 max pooling; 16 channels, 3×3 kernel, ReLU, and 2×2 max pooling; 32 channels, 3×3 kernel, ReLU, and 2×1 max pooling. This is followed by linear layers of sizes 288×64 , 64×32 , and 32×2 , with ReLU activation function in between. The final prediction layer is log softmax. Training involves 100 epochs with a learning rate of 0.0001 and negative log-likelihood loss.

For generating a combination matrix (see Fig. 2a), we initially sample an adjacency matrix following the Erdos-Renyi model with a connection probability of 0.2. Subsequently, we set the combination weights using the averaging rule [39, Chapter 14]. During the inference, we let the central agent be malicious and to observe an object from the opposite class than the rest of the network – see Fig. 2b.

Since we only have 10 objects of each class, having only a handful of objects as a test subset is not enough to provide a reliable accuracy metrics of how well we can identify the

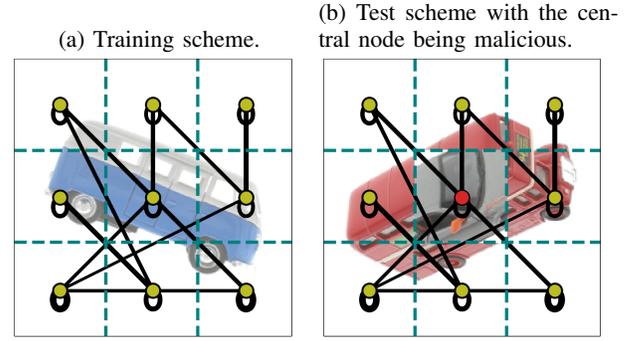


Fig. 2: Observation map of each agent.

(a) Accuracy of the social learning strategy to predict θ_0 . (b) Malicious detection accuracy and learned graph.

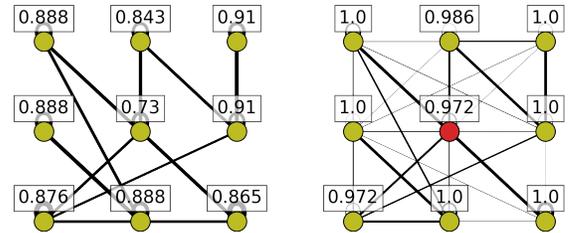


Fig. 3: Accuracy of the adaptive social learning strategy [7] and Algorithm 1. The colors correspond to the output of the algorithms with yellow for θ_0 and red for θ_1 . Per each fold, the accuracy for the social learning is calculated based on average over 100 past iterations.

malicious agent. Thus, we perform a cross-validation procedure where at first, we train the CNNs on 9 objects from each class, leaving 1 object from each class for testing purposes. On average, the cross-validation accuracy of standalone classifiers is **0.68**. The value is relatively low due to a small training set and limited observation available at each agent. Given that many folds had some classifiers with an accuracy below 0.5, we decided to retain only those folds where each agent achieved at least 0.5 accuracy. As a result, we are left with 72 folds instead of 100 with the mean accuracy of standalone classifiers equal to **0.81**.

Finally, we proceed with the adaptive social learning strategy performed at each fold with $\delta = 0.1$. The network observes a moving object from the test set of the class “bus” during 480 iterations (so that, on average, each frame is shown 3 times), whilst the central agents observes an object of a class “car” – see Fig. 2b. We can see that despite the presence of the malicious agent, the average belief of each agent tends towards the correct hypothesis θ_0 (see Fig. 3a) with the mean accuracy **0.8**. However, as depicted in Fig. 3b, the algorithm is able to identify the malicious agent achieving the mean accuracy **0.99**.

REFERENCES

- [1] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.
- [2] X. Zhao and A. H. Sayed, "Learning over social networks via diffusion adaptation," in *Proc. Asilomar Conference on Signals, Systems and Computers*, 2012, pp. 709–713.
- [3] H. Salami, B. Ying, and A. H. Sayed, "Social learning over weakly connected graphs," *IEEE Trans. Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 222–238, 2017.
- [4] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.
- [5] P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie, "Foundations of non-Bayesian social learning," *Columbia Business School Research Paper*, no. 15-95, 2017.
- [6] —, "A theory of non-Bayesian social learning," *Econometrica*, vol. 86, no. 2, pp. 445–490, 2018.
- [7] V. Bordinon, V. Matta, and A. H. Sayed, "Adaptive social learning," *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 6053–6081, 2021.
- [8] —, "Partial information sharing over social learning networks," *IEEE Transactions on Information Theory*, vol. 69, no. 3, pp. 2033–2058, 2023.
- [9] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161–6179, 2018.
- [10] Y. İnan, M. Kayaalp, E. Telatar, and A. H. Sayed, "Social learning under randomized collaborations," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 115–120.
- [11] J. Z. Hare, C. A. Uribe, L. Kaplan, and A. Jadbabaie, "Non-Bayesian social learning with uncertain models," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4178–4193, 2020.
- [12] C. A. Uribe, A. Olshevsky, and A. Nedić, "Nonasymptotic concentration rates in cooperative learning—part i: Variational non-Bayesian social learning," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 3, pp. 1128–1140, 2022.
- [13] D. Gale and S. Kariv, "Bayesian learning in social networks," *Games and Economic Behavior*, vol. 45, no. 2, pp. 329–346, 2003.
- [14] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.
- [15] J. Hkzla, A. Jadbabaie, E. Mossel, and M. A. Rahimian, "Bayesian decision making in groups is hard," *Operations Research*, vol. 69, no. 2, pp. 632–654, 2021.
- [16] M. G. Rabbat and R. D. Nowak, "Decentralized source localization and tracking [wireless sensor networks]," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Montreal, Canada, 2004, pp. 921–924.
- [17] M. Rabbat, R. Nowak, and J. Bucklew, "Robust decentralized source localization via averaging," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, Philadelphia, PA, 2005, pp. 1057–1060.
- [18] P. Hu, V. Bordinon, M. Kayaalp, and A. H. Sayed, "Non-asymptotic performance of social machine learning under limited data," *arXiv:2306.09397*, 2023.
- [19] V. Bordinon, V. Matta, and A. H. Sayed, "Social learning with partial information sharing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 5540–5544.
- [20] V. Shumovskaia, M. Kayaalp, M. Cemri, and A. H. Sayed, "Discovering influencers in opinion formation over social graphs," *IEEE Open Journal of Signal Processing*, vol. 4, pp. 188–207, 2023.
- [21] V. Shumovskaia, M. Kayaalp, and A. H. Sayed, "Distributed decision-making for community structured networks," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, April 2024.
- [22] —, "Social learning in community structured graphs," *arXiv:2312.12186*, 2023.
- [23] A. Mitra, J. A. Richards, and S. Sundaram, "A new approach for distributed hypothesis testing with extensions to byzantine-resilience," in *American Control Conference (ACC)*, Philadelphia, PA, 2019, pp. 261–266.
- [24] L. Su and N. H. Vaidya, "Defending non-bayesian learning against adversarial attacks," *Distributed Computing*, vol. 32, pp. 277–289, 2019.
- [25] H. Zhang, Y. Li, Y. Hu, Y. Chen, and H. V. Zhao, "Measuring the hazard of malicious nodes in information diffusion over social networks," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 476–481.
- [26] H. Zhang, Y. Li, Y. Chen, and H. V. Zhao, "Smart evolution for information diffusion over social networks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1203–1217, 2021.
- [27] V. Krishnamurthy and W. Hoiles, "Afriat's test for detecting malicious agents," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 801–804, 2012.
- [28] C. Zhao, J. He, and J. Chen, "Resilient consensus with mobile detectors against malicious attacks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 60–69, 2018.
- [29] V. P. Illiano and E. C. Lupu, "Detecting malicious data injections in wireless sensor networks: A survey," *ACM Computing Surveys (CSUR)*, vol. 48, no. 2, pp. 1–33, 2015.
- [30] T. Pang, C. Du, Y. Dong, and J. Zhu, "Towards robust detection of adversarial examples," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [31] H. Zhang, M. A. Alim, X. Li, M. T. Thai, and H. T. Nguyen, "Misinformation in online social networks: Detect them all with a limited budget," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 3, pp. 1–24, 2016.
- [32] S. T. Smith, E. K. Kao, E. D. Mackin, D. C. Shah, O. Simek, and D. B. Rubin, "Automatic detection of influential actors in disinformation networks," *Proc. National Academy of Sciences*, vol. 118, no. 4, p. e2011216118, 2021.
- [33] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks," in *NDSS*, vol. 13, 2013, pp. 83–91.
- [34] M. Tomaiuolo, G. Lombardo, M. Mordonini, S. Cagnoni, and A. Poggi, "A survey on troll detection," *Future Internet*, vol. 12, no. 2, p. 31, 2020.
- [35] P. Fornacciarri, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on Twitter," *Computers in Human Behavior*, vol. 89, pp. 258–268, 2018.
- [36] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, "Aggression detection through deep neural model on Twitter," *Future Generation Computer Systems*, vol. 114, pp. 120–129, 2021.
- [37] L. Xing, K. Deng, H. Wu, P. Xie, H. V. Zhao, and F. Gao, "A survey of across social networks user identification," *IEEE Access*, vol. 7, pp. 137 472–137 488, 2019.
- [38] B. Qiu, Y. Li, Y. Chen, and H. V. Zhao, "Controlling information diffusion with irrational users," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 482–485.
- [39] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends® in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014. [Online]. Available: <http://dx.doi.org/10.1561/22000000051>
- [40] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, NY, 2013.
- [41] A. H. Sayed, *Inference and Learning from Data*. Cambridge University Press, 2022, vols. 1–3.
- [42] V. Shumovskaia, K. Ntemos, S. Vlaski, and A. H. Sayed, "Explainability and graph learning from social interactions," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 946–959, 2022.
- [43] A. Kanazaki, Y. Matsushita, and Y. Nishida, "Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5010–5019.