Truth Learning in Social and Adversarial Settings

Júlia Križanová[1] Rhett Olson[2] Filip Úradník[1] Amanda Wang[3]

July 17, 2024

[1]Charles University, Prague, Czech Republic
[2]University of Minnesota, Minneapolis, Minnesota, USA
[3]Princeton University, Princeton, New Jersey, USA

Introduction

Definition (Social learning network)

A social learning network is $\mathcal{N} := (G, q, p)$*, where*

1. $G = (V, E)$ *is a directed graph with agents as vertices,*
2. $q \in (0, 1)$ *is the prior probability of* $\theta = 1$*,*
3. $p \in \left(\frac{1}{2}, 1\right)$ *is the accuracy of agents' private signals* $s_v \in \{0, 1\}$*.*

Introduction

Agents choose actions in some sequence $\sigma \in \Sigma_n$.

When making decisions, agent $v \in V$ has access to

- the private signal s_v,
- actions of neighbors, who chose before v:

$$N_v = \{u \in V \mid uv \in E \wedge \sigma(v) > \sigma(u)\}.$$

When making decisions, agents use an *aggregation rule* μ.

Aggregation Rules

- Bayesian model

$$\mu^B(s_v, N_v) = \begin{cases} 1 & \text{if } \Pr[\theta = 1 \mid s_v, N_v] > \frac{1}{2}, \\ 0 & \text{if } \Pr[\theta = 0 \mid s_v, N_v] > \frac{1}{2}, \\ s_v & \text{otherwise.} \end{cases}$$

Aggregation Rules

- Bayesian model

$$\mu^B(s_v, N_v) = \begin{cases} 1 & \text{if } \Pr[\theta = 1 \mid s_v, N_v] > \frac{1}{2}, \\ 0 & \text{if } \Pr[\theta = 0 \mid s_v, N_v] > \frac{1}{2}, \\ s_v & \text{otherwise.} \end{cases}$$

- Simple majority vote

$$\mu^M(s_v, N_v) = \begin{cases} 1 & \text{if } s_v + \sum_{u \in N_v} a_u > \frac{1}{2}(|N_v| + 1), \\ 0 & \text{if } s_v + \sum_{u \in N_v} a_u < \frac{1}{2}(|N_v| + 1), \\ s_v & \text{otherwise.} \end{cases}$$

Learning Rate

Definition

The expected learning rate of a network $\mathcal{N}$ under the ordering σ and aggregation rule μ is

$$\mathcal{L}(\mathcal{N}, \sigma, \mu) := \frac{1}{n} \sum_{v \in V} \Pr_{\theta, s}[a_v = \theta].$$

NP-hardness

Definition (Network Learning)

Suppose some given aggregation rule μ. NETWORK LEARNING *problem is to decide for a network* $\mathcal{N}$, *and a constant* $\varepsilon \in (0, 1)$ *whether*

$$(\exists \sigma \in \Sigma_n) \quad \mathcal{L}(\mathcal{N}, \sigma, \mu) \geq 1 - \varepsilon.$$

We will focus on the Majority Dynamics setting, meaning $\mu = \mu^M$.

NP-hardness

Definition (Network Learning)

Suppose some given aggregation rule μ. Network Learning *problem is to decide for a network $\mathcal{N}$, and a constant $\varepsilon \in (0,1)$ whether*

$$(\exists \sigma \in \Sigma_n) \quad \mathcal{L}(\mathcal{N}, \sigma, \mu) \geq 1 - \varepsilon.$$

We will focus on the Majority Dynamics setting, meaning $\mu = \mu^M$.

NP-hardness

Definition (Network Learning)

Suppose some given aggregation rule μ. NETWORK LEARNING *problem is to decide for a network $\mathcal{N}$, and a constant $\varepsilon \in (0,1)$ whether*

$$(\exists \sigma \in \Sigma_n) \quad \mathcal{L}(\mathcal{N}, \sigma, \mu) \geq 1 - \varepsilon.$$

We will focus on the Majority Dynamics setting, meaning $\mu = \mu^M$.

NP-hardness

Definition (Network Learning)

Suppose some given aggregation rule μ*.* NETWORK LEARNING *problem is to decide for a network* $\mathcal{N}$*, and a constant* $\varepsilon \in (0,1)$ *whether*

$$(\exists \sigma \in \Sigma_n) \quad \mathcal{L}(\mathcal{N}, \sigma, \mu) \geq 1 - \varepsilon.$$

We will focus on the Majority Dynamics setting, meaning $\mu = \mu^M$.

NP-hardness

Definition (Network Learning)

Suppose some given aggregation rule μ. Network Learning *problem is to decide for a network* $\mathcal{N}$*, and a constant* $\varepsilon \in (0,1)$ *whether*

$$(\exists \sigma \in \Sigma_n) \quad \mathcal{L}(\mathcal{N}, \sigma, \mu) \geq 1 - \varepsilon.$$

We will focus on the Majority Dynamics setting, meaning $\mu = \mu^M$.

NP-hardness

Conjecture

NETWORK LEARNING *is NP-hard.*

NP-hardness

Conjecture

NETWORK LEARNING *is NP-hard.*

- Reduce 3-SAT to NETWORK LEARNING.

NP-hardness

Conjecture

Network Learning *is NP-hard.*

- Reduce 3-SAT to Network Learning.
- *Goal:* Given a formula φ, construct a network $\mathcal{N}$ and choose ε such that the maximal $\mathcal{L}$ exceeds $1 - \varepsilon$ iff φ is satisfiable.

NP-hardness: Proof Intuition

- *Goal:* Given a formula φ, construct $\mathcal{N}$, ε s.t. maximal $\mathcal{L}$ exceeds $1 - \varepsilon$ iff φ is satisfiable.
  $\implies$ Design $\mathcal{N}$ so maximal $\mathcal{L}$ increases with $\#$ satisfied clauses!

NP-hardness: Proof Intuition

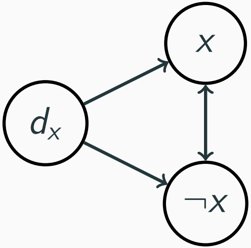
- *Goal:* Given a formula φ, construct $\mathcal{N}$, ε s.t. maximal $\mathcal{L}$ exceeds $1 - \varepsilon$ iff φ is satisfiable.
  $\implies$ Design $\mathcal{N}$ so maximal $\mathcal{L}$ increases with $\#$ satisfied clauses!
- Ordering over *variable gadgets (cells)* encodes boolean variable assignments $x_i = \{T, F\}$.

NP-hardness: Proof Intuition

- *Goal:* Given a formula φ, construct $\mathcal{N}$, ε s.t. maximal $\mathcal{L}$ exceeds $1 - \varepsilon$ iff φ is satisfiable.
  $\implies$ Design $\mathcal{N}$ so maximal $\mathcal{L}$ increases with $\#$ satisfied clauses!
- Ordering over *variable gadgets (cells)* encodes boolean variable assignments $x_i = \{T, F\}$.
- *Clause gadgets* aggregate variables so learning rate is much higher if satisfied.

NP-hardness: Proof Construction

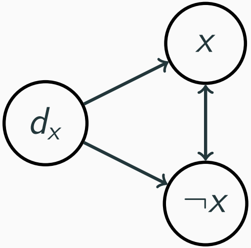
- *Variable cell* for each variable



Cell of a variable x.

NP-hardness: Proof Construction

- *Variable cell* for each variable
- $x = T \quad \leftrightarrow \quad \sigma(x) > \sigma(\neg x)$

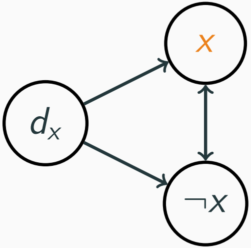


Cell of a variable x.

NP-hardness: Proof Construction

- *Variable cell* for each variable
- $x = T \quad \leftrightarrow \quad \sigma(x) > \sigma(\neg x)$

$\Rightarrow$ Higher probability of being correct

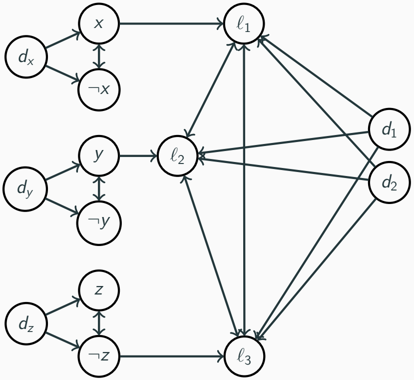


Cell of a variable x.

NP-hardness: Proof Construction

- *Variable cell* for each variable
- $x = T \quad \leftrightarrow \quad \sigma(x) > \sigma(\neg x)$

$\Rightarrow$ Higher probability of being correct



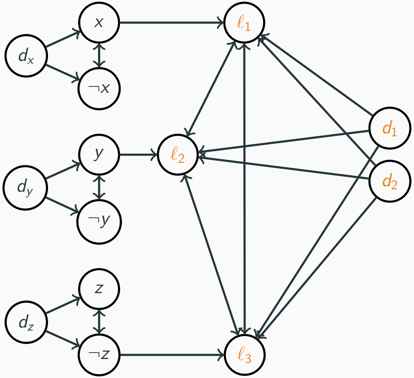
Example for $\varphi = x \vee y \vee \neg z$.

NP-hardness: Proof Construction

- *Variable cell* for each variable
- $x = T \quad \leftrightarrow \quad \sigma(x) > \sigma(\neg x)$

$\Rightarrow$ Higher probability of being correct

- *Clause gadget* for each clause



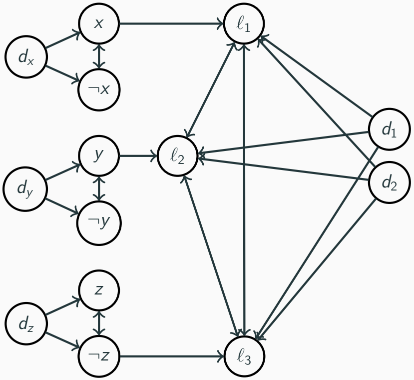
Example for $\varphi = x \vee y \vee \neg z$.

NP-hardness: Proof Construction

- *Variable cell* for each variable
- $x = T \quad \leftrightarrow \quad \sigma(x) > \sigma(\neg x)$

$\Rightarrow$ Higher probability of being correct

- *Clause gadget* for each clause
- Satisfied clause $\Rightarrow$ higher $\mathcal{L}$



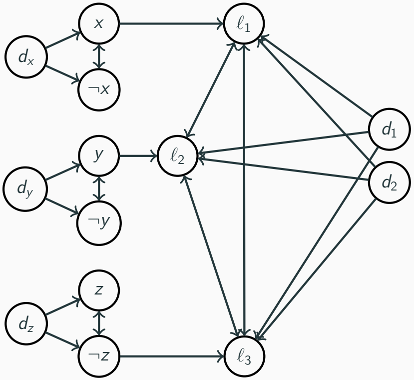
Example for $\varphi = x \vee y \vee \neg z$.

NP-hardness: Proof Construction

- *Variable cell* for each variable
- $x = T \quad \leftrightarrow \quad \sigma(x) > \sigma(\neg x)$

$\Rightarrow$ Higher probability of being correct

- *Clause gadget* for each clause
- Satisfied clause $\Rightarrow$ higher $\mathcal{L}$
- Gap between SAT and non-SAT



Example for $\varphi = x \vee y \vee \neg z$.

NP-hardness

For suitably chosen p, q, ε, it holds that

$$\text{formula } \varphi \text{ is satisfiable} \iff \textsc{Network Learning} \text{ answers yes.}$$

NP-hardness

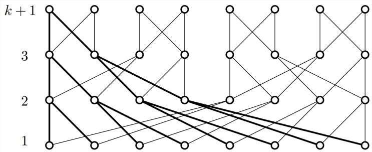
For suitably chosen p, q, ε, it holds that

formula φ is satisfiable $\iff$ $\textsc{Network Learning}$ answers yes.

$\therefore$ $\textsc{Network Learning}$ with μ^M is NP-hard.

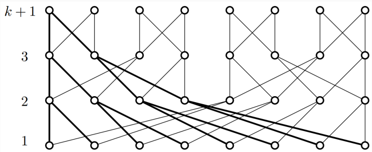
Adversarial Agents

- Some agents might deliberately mislead the group by reporting the opposite of θ.



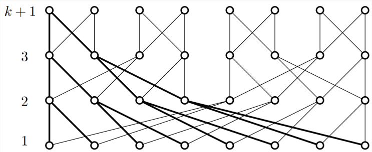
Adversarial Agents

- Some agents might deliberately mislead the group by reporting the opposite of θ.
- We want networks to be robust against such adversaries.



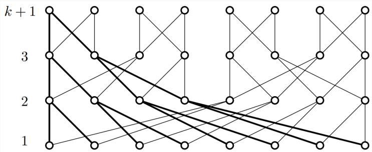
Adversarial Agents

- Some agents might deliberately mislead the group by reporting the opposite of θ.
- We want networks to be robust against such adversaries.
- We study for specific types of networks and configurations of adversaries whether adversaries can affect the learning rate of non-adversaries.



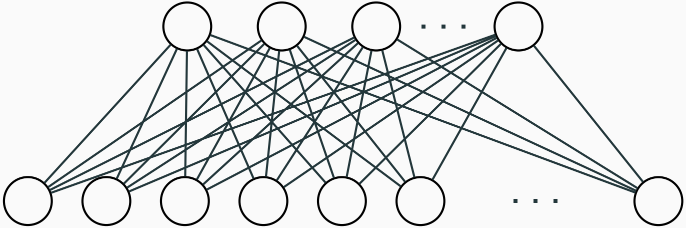
Adversarial Agents

- Some agents might deliberately mislead the group by reporting the opposite of θ.
- We want networks to be robust against such adversaries.
- We study for specific types of networks and configurations of adversaries whether adversaries can affect the learning rate of non-adversaries.
- We studied the Butterfly Network and the Celebrity Network.



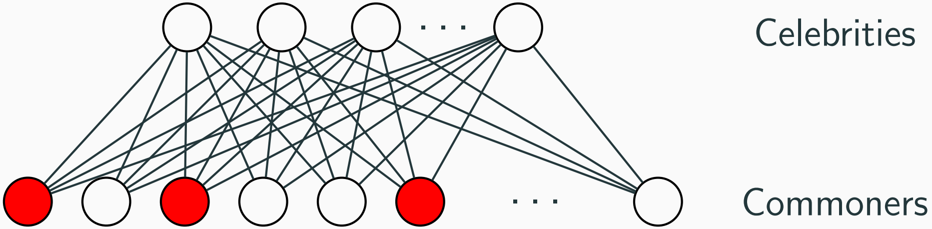
The Celebrity Network

- Complete Bipartite graph with M "celebrities" and N "commoners", with $N >> M$. (Bahar et al., 2020)



The Celebrity Network

- Complete Bipartite graph with M "celebrities" and N "commoners", with $N >> M$. (Bahar et al., 2020)
- We show that it is robust against $\mathcal{O}(N)$ adversarial commoners under a uniformly random decision ordering.



Robustness of Celebrity Network

- Suppose the non-adversarial agents' private signals have probability $0.5 + \delta$ of being correct, for $\delta \in (0, 1/2)$.

Robustness of Celebrity Network

- Suppose the non-adversarial agents' private signals have probability $0.5 + \delta$ of being correct, for $\delta \in (0, 1/2)$.
- Suppose also that there is some $\alpha \in [0, \frac{\delta}{0.5+\delta})$ such that the celebrity network contains αN adversarial commoners.

Robustness of Celebrity Network

- Suppose the non-adversarial agents' private signals have probability $0.5 + \delta$ of being correct, for $\delta \in (0, 1/2)$.
- Suppose also that there is some $\alpha \in [0, \frac{\delta}{0.5+\delta})$ such that the celebrity network contains αN adversarial commoners.
- Suppose the decision order is uniformly random.

Robustness of Celebrity Network

- Suppose the non-adversarial agents' private signals have probability $0.5+\delta$ of being correct, for $\delta \in (0, 1/2)$.
- Suppose also that there is some $\alpha \in [0, \frac{\delta}{0.5+\delta})$ such that the celebrity network contains αN adversarial commoners.
- Suppose the decision order is uniformly random.
- Given any $\epsilon > 0$, the expected learning rate for the network is at least $1-\alpha-\epsilon$ for sufficiently large networks.

Robustness of Celebrity Network

Proof Sketch:

Robustness of Celebrity Network

Proof Sketch:

- Under random ordering, the first celebrity will observe a large pool of commoners WHP (e.g. probability $> 1 - \frac{\epsilon}{8}$).

Robustness of Celebrity Network

Proof Sketch:

- Under random ordering, the first celebrity will observe a large pool of commoners WHP (e.g. probability $> 1 - \frac{\epsilon}{8}$).
- A majority of these commoners will correctly predict the ground truth θ WHP (this can be shown using Chebyshev's inequality).

Robustness of Celebrity Network

Proof Sketch:

- Under random ordering, the first celebrity will observe a large pool of commoners WHP (e.g. probability $> 1 - \frac{\epsilon}{8}$).
- A majority of these commoners will correctly predict the ground truth θ WHP (this can be shown using Chebyshev's inequality).
- The first celebrity will mimic this majority.
- All non-adversaries after this first celebrity will mimic the action of the first celebrity.

References i

- Bahar, Gal et al. (2020). **"Multi-issue social learning".** In: *Mathematical Social Sciences* 104, pp. 29–39.
- Easley, David, Jon Kleinberg, et al. (2010). **Networks, crowds, and markets: Reasoning about a highly connected world.** Vol. 1. Cambridge university press Cambridge.
- Hazła, Jan et al. (2019). **"Reasoning in Bayesian opinion exchange networks is PSPACE-hard".** In: *Conference on Learning Theory*. PMLR, pp. 1614–1648.

References ii

- Lu, Kevin et al. (2024). **Enabling Asymptotic Truth Learning in a Social Network.** Manuscript submitted for publication.
- Mossel, Elchanan, Joe Neeman, and Omer Tamuz (2013). **"Majority dynamics and aggregation of information in social networks".** In: *Autonomous Agents and Multi-Agent Systems* 28, 408–429.

Acknowledgments

This work was made possible by the Rutgers DIMACS REU program. Thank you as well to Professor Jie Gao and Ph.D. student Kevin Lu for their help and leadership.

Thank you to the National Science Foundation for funding this project through the grant CNS-2150186 and the REU supplement to NSF 2208663 -Collaborative Research: AF: Small: Promoting Social Learning Amid Interference in the Age of Social Media.

Filip Úradník and Júlia Križanová were supported by CoSP, a project funded by European Union's Horizon 2020 research and innovation programme, grant agreement No. 823748.