Local non-Bayesian social learning with stubborn agents

Daniel Vial, Vijay Subramanian

ECE Department, University of Michigan
Motivation

Social learning in the presence of malicious agents

Most prominent example: fake news on social networks

[Shearer, Gottfried 2017] [Shearer 2018] [Allcott, Gentzkow 2017]
Overview

Salient features:

1. Simultaneous consumption/discussion of news
2. Legitimate news partially reveals “truth”
3. Fake news more likely in “echo chambers”

We analyze model incorporating these features:

1. Agents receive signals/share beliefs about true state $\theta$
2. Regular agents: signals = noisy observations of $\theta$
3. Stubborn agents: signals uncorrelated with $\theta$; ignore others’ beliefs

Main questions:

- Do stubborn agents prevent regular agents from learning $\theta$?
- How can stubborn agents maximize influence?
Learning model (basic ingredients)

True state $\theta \in (0, 1)$, (regular) agents $A$, stubborn agents/bots $B$

Signals at time $t$: $s_t(i) \sim \text{Bernoulli}(\theta)$ for $i \in A$, $s_t(i) = 0$ for $i \in B$

Beliefs at time $t$: $\text{Beta}(\alpha_t(i), \beta_t(i))$ for $i \in A \cup B$

If $j \to i$ in graph, $i$ observes $\alpha_{t-1}(j), \beta_{t-1}(j)$ at $t$; $i \in B$ has only self-loop
Learning model (belief updates)

How should $i$ use signal $s_t(i) + \text{neighbor parameters } \{\alpha_{t-1}(j), \beta_{t-1}(j) : j \rightarrow i\}$?

We adopt non-Bayesian model similar to [Jadbabaie et al. 2012]

**Bayesian update using signal, then average with neighbors in graph:**

$$\alpha_t(i) = (1 - \eta)(\alpha_{t-1}(i) + s_t(i)) + \frac{\eta}{d_{in}(i)} \sum_{j \in A \cup B : j \rightarrow i} \alpha_{t-1}(j)$$

$$\beta_t(i) = (1 - \eta)(\beta_{t-1}(i) + 1 - s_t(i)) + \frac{\eta}{d_{in}(i)} \sum_{j \in A \cup B : j \rightarrow i} \beta_{t-1}(j)$$

**Quantity of interest:**

$$\theta_t(i) = \mathbb{E} [\text{Beta}(\alpha_t(i), \beta_t(i))] = \frac{\alpha_t(i)}{\alpha_t(i) + \beta_t(i)}$$

(View as summary statistic of $i$'s belief/opinion at $t$)
As learning horizon (i.e. number belief updates) grows . . .

- . . . agents receive more unbiased observations
- . . . influence of bots spreads

Learning horizon plays important, but non-obvious role

Difficult to analyze finite horizon for fixed graph

- Will consider sequence $\{G_n\}_{n \in \mathbb{N}}$ of random graphs, where $G_n$ has $n$ agents
- Will consider horizon $T_n \in \mathbb{N}$ for $G_n$ (finite for each finite $n$)
Graph model

1. Realize \( \{d_{out}(i), d_{in}^A(i), d_{in}^B(i)\}_{i=1}^{n} \) satisfying

\[
d_{out}(i) \in \mathbb{N}, \ d_{in}^A(i) \in \mathbb{N}, \ d_{in}^B(i) \in \mathbb{Z}^+, \ \sum_{i=1}^{n} d_{out}(i) = \sum_{i=1}^{n} d_{in}^A(i) \ a.s.
\]

2. From \( \{d_{out}(i), d_{in}^A(i)\}_{i=1}^{n} \), construct sub-graph with nodes \( A = \{1, \ldots, n\} \) via directed configuration model [Chen, Olvera-Cravioto 2013]

3. Connect \( d_{in}^B(i) \) bots (with only self-loop) to each \( i \in A \)

Here bot connections \( \{d_{in}^B(i)\}_{i=1}^{n} \) given; later, will consider optimal connections.
Key random variable: “density” of (regular) agents, measured as

\[
\tilde{p}_n = \sum_{i=1}^{n} \frac{d^A_{\text{in}}(i)}{d^A_{\text{in}}(i) + d^B_{\text{in}}(i)} \times \frac{d_{\text{out}}(i)}{\sum_{j=1}^{n} d_{\text{out}}(j)}
\]

Fraction in-neighbors trying to learn
Sample w.r.t. out-degree distribution

Assumption 1 (for belief convergence):
- \(\lim_{n \to \infty} \mathbb{P}(|\tilde{p}_n - p_n| > \delta_n) = 0\) for some \(\{p_n\}_{n \in \mathbb{N}}, \{\delta_n\}_{n \in \mathbb{N}} \subset (0, 1)\) s.t. \(\lim_{n \to \infty} \delta_n = 0\)
- \(\lim_{n \to \infty} T_n = \infty\)

Assumption 2 (for branching process approximation):
- Sparse degrees (finite mean/variance) with high probability
- \(T_n = O(\log n)\)
\(\Rightarrow\) Guarantees \(\theta_{T_n}(i)\) depends on \(o(n)\) other agents ("local" learning)
Main result

**Theorem**

Given assumptions, we have for $i^* \sim \{1, \ldots, n\}$ uniformly,

$$
\theta T_n(i^*) \xrightarrow{\mathbb{P}_{n \to \infty}} \begin{cases} 
\theta, & T_n(1 - p_n) \xrightarrow{n \to \infty} 0 \\
\theta (1 - e^{-Kn})/(K\eta), & T_n(1 - p_n) \xrightarrow{n \to \infty} K \in (0, \infty) \\
0, & T_n(1 - p_n) \xrightarrow{n \to \infty} \infty
\end{cases}
$$

Illustration, assuming $T_n, p_n$ related as $T_n \propto (1 - p_n)^{-C}$:

$$
\lim_{n \to \infty} \theta T_n(i^*)
$$

0

$T_n = (1 - p_n)^{-C}, C \in (0,1)$  $T_n = K(1 - p_n)^{-1}, K \in (0,\infty)$  $T_n = (1 - p_n)^{-C}, C \in (1,\infty)$
Remarks on main result

Again assuming $T_n, p_n$ related as $T_n \propto (1 - p_n)^{-C}$:

1. *Phase transition* occurs (small change to $C \approx 1 \Rightarrow$ big change belief)
2. For fixed $p_n$, agents initially (at small $T_n$) learn, later (at large $T_n$) forget!
3. For fixed $T_n \propto (1 - p_n)^{-1}$, bots experience “diminishing returns”
4. When $T_n(1 - p_n) \rightarrow K \in (0, \infty)$, limiting belief $= \theta(1 - e^{-K\eta})/(K\eta)$:
   - As $\eta \rightarrow 0$, agents ignore network, belief $\rightarrow \theta$
   - As $\eta \rightarrow 1$, belief $\rightarrow \theta(1 - e^{-K})/K$ (not $\rightarrow 0$, “discontinuity”)

\[
\lim_{n \to \infty} T_n(i^n) = \theta
\]
If $p_n \to p < 1$ (i.e. bots non-vanishing), stronger result holds:

**Theorem**

Suppose $p_n \to p \in (0, 1)$, so that $\theta_{T_n}(i^*) \to 0$ in $\mathbb{P}$. Then, under slightly stronger assumptions, and for any $\epsilon > 0$,

$$|\{i \in A : \theta_{T_n}(i) > \epsilon\}| = o(n) \text{ with high probability as } n \to \infty.$$  

“Slightly stronger assumptions”:

- $T_n = \Omega(\log n)$ (instead of just $T_n \to \infty$)
- Minimum rates of convergence for “with high probability” statements
Key ideas of proof (1/2)

Recall parameter updates:

$$\alpha_t(i) = (1 - \eta)(\alpha_{t-1}(i) + s_t(i)) + \frac{\eta}{d_{in}(i)} \sum_{j \in A \cup B: j \rightarrow i} \alpha_{t-1}(j)$$  \hspace{1cm} (1)$$

$$\beta_t(i) = (1 - \eta)(\beta_{t-1}(i) + 1 - s_t(i)) + \frac{\eta}{d_{in}(i)} \sum_{j \in A \cup B: j \rightarrow i} \beta_{t-1}(j)$$  \hspace{1cm} (2)$$

Assume $\alpha_0(j) = \beta_0(j) = o(T_n)$ $\forall j$ and define

- $P =$ column-normalized adjacency matrix
- $e_i =$ unit vector in $i$-th direction

Then iterating (1)-(2) yields

$$\theta_{T_n}(i) = \frac{1}{T_n} \sum_{\tau=0}^{t-1} s_{t-\tau} (\eta P + (1 - \eta)I)^\tau e_i + o(1)$$

**Interpretation:** take Uniform($\{1, \ldots, T_n\}$)-length lazy random walk from $i$, sample signal of node reached
Key ideas of proof (2/2)

Previous slide: interpret $\theta_{T_n}(i)$ in terms of lazy random walk (LRW)

Bots are absorbing states on this LRW (owing to self-loops)

To analyze beliefs, analyze absorption probabilities

LRW and breadth-first-search graph construction can be done simultaneously

By $T_n = O(\log n)$ and sparsity, LRW explores tree-like sub-graph before horizon

Reduces random process on random graph to much simpler process
(simultaneous construction of tree / computation of absorption probabilities)
Formulation

Previously assumed \( \{d_{out}(i), d_{in}^A(i), d_{in}^B(i)\}_{i=1}^n \) given

Now suppose \( \{d_{out}(i), d_{in}^A(i)\}_{i=1}^n \) given, adversary chooses \( \{d_{in}^B(i)\}_{i=1}^n \)

By main result, adversary (with budget \( b \in \mathbb{N} \)) should solve

\[
\min_{\{d_{in}^B(i)\}_{i=1}^n \in \mathbb{Z}_+^n} \sum_{i=1}^n \frac{d_{in}^A(i)}{d_{in}^A(i) + d_{in}^B(i)} \sum_{j=1}^n d_{out}(j) s.t. \sum_{i=1}^n d_{in}^B(i) \leq b
\]

Key random variable \( \tilde{p}_{n} \) shown previously

Integer program (IP), so we devise approximation scheme
Approximation scheme

Independently attach each bot to $i$-th agent with probability proportional to

$$\max \left\{ d_{in}^A(i) \left( \sqrt{\lambda^* \frac{d_{out}(i)}{d_{in}^A(i)}} - 1 \right), 0 \right\}$$  \hspace{1cm} (3)

- (3) is solution to LP relaxation of IP; $\lambda^* > 0$ is efficiently computable
- Intuition: bots want to connect to $i$-th agent only if $\frac{d_{out}(i)}{d_{in}^A(i)} \geq \frac{1}{\lambda^*}$, i.e. only if $i$ is influential ($d_{out}(i)$ large) + susceptible to influence ($d_{in}^A(i)$ small)

Theorem

For any $\delta > 0$, scheme gives $(2 + \delta)$-approximation with high probability, i.e.

$$\lim_{n \to \infty} P \left( \frac{\text{objective for approximation scheme}}{\text{objective for optimal scheme}} > 2 + \delta \right) = 0.$$
Empirical performance

For real social networks, our approximation scheme outperforms heuristics, even those using network structure.

(Networks from [SNAP Datasets: Stanford Large Network Dataset Collection])

Ultimately, new insights into vulnerabilities of social networks
Most similar models in literature

[Azzimonti, Fernandes 2018]
- (Almost) same belief update (minor differences to bot behavior)
- Only empirical results (allows for richer model, e.g. time-varying graph)

[Jadbabaie et al. 2012]
- Communicate distributions, not parameters, i.e.

\[ \mu_t(i) = \eta_{ii} \text{BU}(\mu_{t-1}(i), s_t(i)) + \sum_{j \neq i} \eta_{ji} \mu_{t-1}(j) \]

where \( \mu \) terms are distributions, \( \sum_j \eta_{ji} = 1 \), \( \text{BU} = \) “Bayesian update”

- Richer belief update, but stronger assumptions:
  1. Fixed, strongly-connected graph
  2. Infinite horizon
  3. No stubborn agents
Other relevant works

View our model as perturbation of classical deGroot model [DeGroot 1974]:

$$\theta_t = \theta_{t-1} W$$ where $\theta_t, \theta_{t-1} \in \mathbb{R}^n$ and $W$ is column-stochastic

Extensively studied, see surveys [Acemoglu, Ozdaglar 2011; Golub, Sadler 2017]

[Rahimian, Shahrampour, Jadabaie 2015]: adopt belief of random neighbor, also relates to random walk (but need strong connectedness + infinite horizon)

[Acemoglu, Ozdaglar, ParandehGheibi 2010]: “forceful” but not fully-stubborn agents $\Rightarrow$ no absorbing states $\Rightarrow$ can use stationarity distribution

Stubborn agents have been considered in consensus setting, but infinite horizon typically assumed, e.g. [Acemoglu et al. 2011; Ghaderi, Srikant 2014]


