Reputation and Manipulation in Social Networks

James Siderius[1], Mohamed Mostagir[2], and Asu Ozdaglar[1]

MIT Economics Theory Lunch

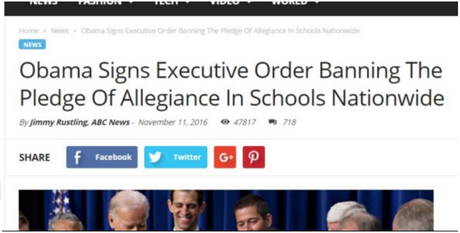
March 5, 2019

[1] MIT EECS LIDS

[2] University of Michigan Ross School of Business

Motivation

- Reputation: Players come to expect certain behavior from their opponent. A principal builds its reputation to sustain long-run relationships with short-lived agents.
- Social learning: Agents try to learn about a state of the world through: (i) direct experience and (ii) communication with others.
- Examples:
    - Fake news: costly to report interesting stories which are truthful, but want continued patronage.
    - Yelp reviewers: provide high-profile critics good restaurant service, work less hard for tourists.



NEWS
FASHION
TECH
VIDEO
WORLD
Home › News › Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide
NEWS
SHARE
Facebook
Twitter
G+

Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide

By Jimmy Rustling, ABC News - November 11, 2016 47817 718

- What environments lead to manipulation?

This Talk

- A framework to study principal-agent(s) interaction in a heterogenous setting.
  - Model combines agents with different sophistication: Bayesian (B) and DeGroot (D).
- Reputation effects are not isolated: people communicate their experiences and beliefs.
  - Principal's problem is more nuanced: network externalities from actions on the beliefs of other agents.
- Can the principal sculpt learning to his advantage?

Related Literature

- Reputation effects
  - Kreps and Wilson (1982), Milgrom and Roberts (1982), Fudenberg-Levine (1989), Gossner (2011)
- Bayesian learning
  - Acemoglu et al (2011), Bikhchandani et al (1992)
- DeGroot-style learning
  - Golub and Jackson (2010), Jadbabaie et al (2012)
- Mixed-learning environments
  - Mueller-Frank (2014), Chandrasekhar et al (2015), Pennycook and Rand (2018)
- Propagation of fake news
  - Candogan and Drakopoulos (2017), Papanastasiou (2018)

Organization of the Paper

- General Framework: Single patient principal playing against n agents, arranged in a social network.
  - Agents have different reasoning abilities: DeGroot or Bayesian.
  - Formal definition of manipulation, study long-term beliefs and asymptotic learning.
  - General results can be used to study any interactions of this form.
- Social Contract Game: Consider a canonical reputation game.
  - How does the distribution of agent types and network structure affect whether the principal can manipulate?
  - Consider both deterministic and random network topologies, including the effects of homophily on manipulation.

General Game

- Time is discrete: $t = 1, 2, \ldots$
- Each agent i has a set of actions $\mathcal{A}_i^a$ and plays some strategy from $S_i^a = \Delta(\mathcal{A}_i^a)$ at every time t.
- Principal has a finite set of actions $\mathcal{A}_i^p$ and must play a time-invariant strategy from $S_i^p = \Delta(\mathcal{A}_i^p)$ for every agent i.
- Payoff of the agent is $u_i^a : \mathcal{A}_i^p \times \mathcal{A}_i^a \to \mathbb{R}$ and payoff for the principal is $u_i^p : \mathcal{A}_i^p \times \mathcal{A}_i^a \to \mathbb{R}$ for the principal, with $u^p = \sum_{i=1}^n u_i^p$.
- Each agent i observes perfectly the action taken by the principal at i (i.e., $a_{i,t}^p$) but does not observe any other actions.

Reputation Setup

- Finite set of principal types $\Omega = \{\tilde{\omega}\} \cup \hat{\Omega}$.
    - $\tilde{\omega}$: a strategic principal
    - $\hat{\omega} \in \hat{\Omega}$: a commitment type which plays some mixed strategy from $S^p \equiv \bigtimes_{i=1}^{n} S_i^p$ for all t.
    - $\mu \in \Delta(\Omega)$: initial distribution over the type of the principal; full support over commitment types.
- Bayesian agents are aware that the principal may be strategic.
- DeGroot agents are mechanical agents who do not anticipate the principal being strategic.
    - Assume there is another commitment type $\hat{\omega}_1$, which may or may not be the strategy played by the strategic principal in equilibrium.
    - Call this a DeGroot conjecture about the play of the "strategic" type.

Learning Process

- DeGroot agents: Beliefs about the type of the principal evolve using common learning heuristic, plus a personal-experience term:

  $$\pi_{i,t+1} = \theta_i g_i(a^p_{i,t}, \pi_{i,t}) + \sum_{i=1}^{n} \alpha_{ij}\pi_{j,t}$$

  where $\theta_i + \sum_{j=1}^{n} \alpha_{ij} = 1$.

  - $\pi_{i,t}$: the belief of agent i at time t.
  - $\theta_i \in (0,1)$ - weight on one's own experience.
  - $g_i : (\mathcal{A}^p_i, \Delta(\Omega_D)) \to \Delta(\Omega_D)$ - function mapping the action of the principal to a belief of the type.
  - α_{ij}: the influence of agent j's opinion on agent i.

- Bayesian agents: neighborhood $\mathcal{N}_i$ of agents in the network. Observe both the private history of play at oneself, $H_{i,t-1}$, and the beliefs in the neighborhood for all $\tau \leqslant t-1$, $\Pi_{i,t}$.
  - Belief updated via Bayes' rule conditioning on both $H_{i,t-1}$ and $\Pi_{i,t}$.
  - Assume that Bayesians report beliefs truthfully.

Information Structure

- Principal knows:
  - Network structure
  - Who is DeGroot and who is Bayesian
  - Actions played by all the agents in the network at each time t (maybe)
- Bayesian agent i knows:
  - Network structure
  - Who is DeGroot and who is Bayesian
  - Action played by the principal at i at every time t
  - Beliefs in neighborhood at every time t
- DeGroot agent i knows:
  - Action played by the principal at i at every time t
  - Beliefs in neighborhood (i.e., $\{j : \alpha_{ij} > 0\}$) at every time t

Strategies

- Principal and agent each choose strategies as follows:
  - Principal: A map $\sigma^p : \Omega \to S^p$ such that $\sigma^p(\hat{\omega}) = \hat{\omega}$ for all $\hat{\omega}$ (i.e., commitment types required to play committed strategy).
  - DeGroot agents: A map $\sigma^a_{i,t} : \Delta(\Omega_D) \to S^a_i$.
  - Bayesian agents: A map $\sigma^a_{i,t} : H_{i,t-1} \times \Pi_{i,t} \to (S^a_i, \pi_{i,t})$.
- The principal maximizes the discounted payoff for $0 < \delta < 1$:

$$\pi^p(\sigma) = (1-\delta)\mathbb{E}\sum_{t=0}^{\infty} \delta^t u^p(a^p_t, a^a_t)$$

as $\delta \to 1$.

- Each agent is myopic:

$$\pi^a_i(\sigma) = \mathbb{E}u^a_i(a^p_{i,t}, a^a_{i,t})$$

An Illustration

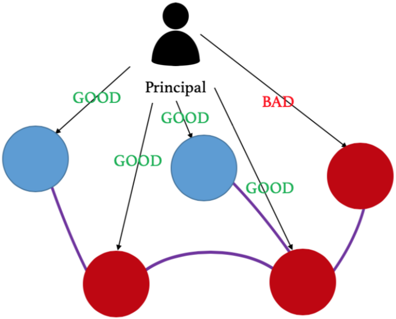


Figure 1: A simple illustration of the reputation game, where the principal elects to either give good or poor restaurant service to each agent in the network (Blue = Bayesian, Red = DeGroot). DeGroot agents form opinions based on their own experience and what their friends think. Bayesian agents form opinions by trying to infer the actions of the principal across the entire network as opposed to just in their immediate neighborhood. Some agents may continue to get treated poorly because they listen to opinions which contradict their experience.

Manipulation

- For agent i, let $BR_i(\sigma^p)$ be the set of best-response strategies $\sigma^a_{i,t}$ to σ^p.
- Agent i is manipulated at time t if the following two conditions hold:
  1. Not a Best-Response: $\sigma^a_{i,t} \notin BR_i(\sigma^p_i(\omega))$.
  2. Principal Benefits: $u^p_i(\sigma^a_{i,t}, \sigma^p(\omega)) > \sup_{\bar{\sigma}^a_{i,t} \in BR_i(\sigma^p_i(\omega))} u^p_i(\bar{\sigma}^a_{i,t}, \sigma^p(\omega))$
- Agent is manipulated in the network game if she is manipulated at times $\{t_\tau\}_{\tau=1}^{\infty}$ for some unbounded sequence.

Definition. (Manipulation)

We say that the network is *susceptible* to manipulation if in some equilibrium of the network game, for all DeGroot conjectures $(\hat{\omega}_1)$, there exists an agent who is manipulated. We say the network is *impervious* to manipulation if in every equilibrium of the network game, there exist DeGroot conjectures such that no agent is manipulated.

Manipulation: Bayesians and DeGroots

- The network of all Bayesian agents: with "well-behaved" payoffs and best-responses, there is no manipulation. Equivalent to n isolated reputation games with a few caveats:
  - Asymptotic learning is the same, but short-run learning dynamics are different with network.
  - Larger set of equilibria with network: principal can be held below sum of lower repuation payoffs.
- The network of all DeGroot agents: generally, there is always manipulation except when θ_i is large and g_i is relatively sophisticated.
  - For example, if $\theta_i = 1$ and g_i is a Bayesian update, then the agent will learn the true type of the principal and play a best-response.
- Most interesting case: mixed-learning environment.

Asymptotic Bayesian Learning

- For any $\sigma^p(\tilde{\omega})$ and type space Ω, in equilibrium, consider the following learning properties (*):
  1. If $\sigma^p(\tilde{\omega}) \neq \hat{\omega}$ for any $\hat{\omega}$, then for all Bayesian agents i, $\lim_{t\to\infty} \pi_{i,t}(\tilde{\omega}) \overset{\text{a.s.}}{\to} 1$.
  2. If $\sigma^p(\tilde{\omega}) = \hat{\omega}$ for some $\hat{\omega}$, then for all Bayesian agents i, $\lim_{t\to\infty} \pi_{i,t}(\tilde{\omega}) + \pi_{i,t}(\hat{\omega}) \overset{\text{a.s.}}{\to} 1$.

Theorem 1. (Bayesian Learning)

Under some regularity conditions, for generic networks $\mathbf{A} \equiv \{\alpha_{ij}\}_{i,j=1}^{n}$ and experience-functions $\mathbf{g}$, the asymptotic learning properties (*) hold.

- Even if $\sigma_i^p = \hat{\omega}_i$ for every Bayesian agent i (i.e., the Bayesians cannot differentiate the types based on on their observations), the Bayesians can deduce the true type of the principal by communicating with the DeGroot agents.

Asymptotic DeGroot Learning

- Assume g_i maps actions into beliefs. Then write the evolution of beliefs as:

$$\boldsymbol{\pi}_{t+1}(\omega) = \mathbf{A}\boldsymbol{\pi}_t(\omega) + \mathbf{g}(\sigma^p(\omega)) \otimes \boldsymbol{\theta}$$

where the matrix $\mathbf{A}$ is given by:

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{D,B} & \mathbf{A}_{D,D} \end{pmatrix}$$

- Let ω^* be the true type of the principal and $\pi_\infty^B(\omega|\omega^*)$ be the belief of type ω conditional on true type ω^*. Then:

$$\mathbf{g}(\sigma^p(\omega)) \sim \left(\pi_\infty^B(\omega|\omega^*)\mathbf{1}'_{|B|}, g_1(a_1^p), g_2(a_2^p), \ldots, g_{|D|}(a_{|D|}^p)\right)'$$
$$\boldsymbol{\theta} = (\mathbf{1}_{|B|}, \theta_1, \theta_2, \ldots, \theta_{|D|})'$$

- Bayesian agents act as stubborn agents who “know” the true type of the principal.
    - Related to literature: Acemoglu et al (2013), Yildiz et al (2013)

Asymptotic DeGroot Learning, cont.

- Assume that we restrict the possible strategies of the principal to pure strategies:
    1. Equilibria easiest to analyze: how are beliefs of principal type confounded by the learning process?
    2. Even when the principal must commit upfront (i.e., invest in a technology and use this throughout the entire horizon), manipulation may still be possible.

Proposition 1. (DeGroot Learning)

Under certain regularity conditions, if σ^p is a pure strategy, as $t \to \infty$, the beliefs of the DeGroot agents about the type of the principal converge almost surely:

$$\boldsymbol{\pi}_{D,\infty} \overset{\text{a.s.}}{\to} (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{g}(\sigma^p) \otimes \boldsymbol{\theta})$$

- Agents are learning a state which can be strategically set by a principal. Principal knows how limit beliefs are induced by the strategy σ^p employed.

Social Contract Game

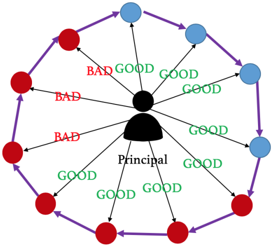
		Agent	
		Opt In	**Opt Out**
Principal	**Good**	$1, 1$	$0, 0$
	Bad	$1 + \epsilon, -\epsilon$	$0, 0$

Table 1: Social Contract Game.

- Static NE: Agent plays **Opt Out** w.p. 1 and principal plays **Bad** w.p. at least $1/(1+\epsilon)$. Pareto sub-optimal.
- Single commitment type $\hat{\omega}$ which plays **Good** always; prior probability $\zeta > 0$. Type space of the principal $(\hat{\omega}, \tilde{\omega})$.
- Reputation equilibria: *Grim Trigger*, *Lagged Best-Response*, etc.
- Experience function $g_i(\mathbf{Good}) = (1, 0)$ and $g_i(\mathbf{Bad}) = (0, 1)$ for all DeGroots.
- Characterization of equilibria as $\zeta \to 0$.

Example: The Ring Network

Assume the first $m = o(n)$ agents along the ring are Bayesian; everyone else is DeGroot with $\theta_i = 1/(n+1)$.



Proposition 2. (Ring Manipulation)

There exists a number D such that for any $0 < \epsilon < (e-1)$, the ring network with $d > D$ DeGroot agents and any number of Bayesian agents is susceptible, and the fraction of manipulated agents is no less than $1 - \log(1 + e\epsilon/(1+\epsilon))$.

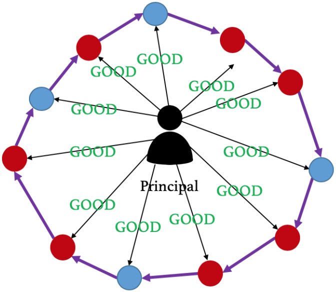
Proof Idea

- Consider a heuristic for the principal's strategy: play **Good** for τ DeGroot agents closest to the Bayesians, then switch to **Bad** for the remaining DeGroots.
  - This gives a lower bound on the number of manipulated agents in equilibrium; no claim that this heuristic is optimal!
  - Heuristic Idea: Those closest to the Bayesians have stronger beliefs about the principal's true type, whereas those farther away tend to believe commitment type because positive experiences drown out Bayesian opinions.
- Need $\pi_{i,\infty}(\hat{\omega}) > \epsilon/(1+\epsilon)$ to ensure that switching to **Bad** does not induce the agent to switch to **Opt Out**:

$$\tau^*(n,\epsilon) = \inf\left\{\tau : \left[\frac{n}{n+1} - \left(\frac{n}{n+1}\right)^{\tau}\right] \cdot \left(\frac{n}{n+1}\right)^{n-m-\tau} > \frac{\epsilon}{1+\epsilon}\right\}$$

which for n large, $\tau^*(n,\epsilon) \approx n\log(1 + e\epsilon/(1+\epsilon))$.

Imperviousness in the Ring



Proposition 2. (Ring Imperviousness)

For any $0 < \epsilon < (e-1)$, in the ring $\Omega(n)$ optimally-placed Bayesians are needed for imperviousness.

- A linear number of Bayesian agents are needed *and* their location is of critical importance.

DeGroot Centrality

- How do we characterize imperviousness in general networks?
  - Belief of a given DeGroot agent is related to its centrality to other DeGroot agents who are having positive experiences.
  - Generalize Katz-Bonacich centrality to account for node-dependent discount factors which depend on the principal's strategy σ^p.
- For slack parameter γ, define the characteristic vector to be:

$$\boldsymbol{\xi}(\gamma) = \left(\begin{array}{c} \mathbf{0}_m \\ \boldsymbol{\theta} \otimes \gamma \end{array} \right)$$

- Then, we have DeGroot centrality given by:

$$\mathcal{D}(\gamma) \equiv (\mathbf{I} - \mathbf{A})^{-1} \boldsymbol{\xi}(\gamma)$$

- For $\zeta \to 0$, DeGroot centrality is equal to $\pi_{i,\infty}(\hat{\omega})$ for DeGroot agents.

Weighted Walk Interpretation

- Most useful to think about DeGroot centrality in terms of weighted walks:
  - Define the weight of a walk $W = i \to v_1 \to v_2 \to \ldots \to v_n \to j$ to be:

$$w_W = \prod_{v_i \to v_{i+1}} \alpha_{v_i, v_{i+1}}$$

  - If $\mathcal{W}_{ij}$ is the set of all walks between i and j not passing through a Bayesian, then:

$$[(\mathbf{I} - \mathbf{A})^{-1}]_{ij} = \sum_{W \in \mathcal{W}_{ij}} w_W < \infty$$

- Main proof technique: bound weighted walks in order to bound DeGroot centrality instead of matrix inversion (only possible in symmetric networks e.g., ring or complete).
- Setting $\boldsymbol{\theta} = (1 - \beta)\mathbf{1}$ and $\boldsymbol{\gamma} = \mathbf{1}$ recovers β-Bonacich centrality, so DeGroot centrality is a generalization.

The General Principal's Problem

- Let us denote by $\boldsymbol{\chi}(\boldsymbol{\gamma})$:

$$\chi_i(\boldsymbol{\gamma}) = 1 - \left(1 - 1_{\mathcal{D}_i(\boldsymbol{\gamma}) > \epsilon/(1+\epsilon)}\right)\left(1 - 1_{\mathcal{D}_i(\boldsymbol{\gamma}) = \mathcal{D}_i(\mathbf{1})}\right)$$

Theorem 2.

For any $\epsilon > 0$, the principal solves:

$$\boldsymbol{\Gamma}^* = \arg\max_{\gamma \in \{0,1\}^{|D|}} \sum_{i \in D} [1 + \epsilon(1 - \gamma_i)]\chi_i(\boldsymbol{\gamma})$$

The network is impervious if $\mathbf{1} \in \boldsymbol{\Gamma}^*$; otherwise it is susceptible.

- Some easy sufficiency conditions from Theorem 2:
  - Imperviousness: $\mathcal{D}_i(\mathbf{1}_{-i}) < \epsilon/(1+\epsilon)$ for every DeGroot i.
  - Susceptible: there exists DeGroot i such that $\mathcal{D}_j(\mathbf{1}_{-i}) > \epsilon/(1+\epsilon)$ for all DeGroots j (including i).

Personal Experience and Societal Norms

- Network Preservation: Form $(\mathbf{A}', \boldsymbol{\theta}')$ from $(\mathbf{A}, \boldsymbol{\theta})$ without changing network structure: $\alpha'_{ij} = \alpha_{ij}(1-\theta'_i)/(1-\theta_i)$ for all DeGroot agents i.
  - Can interpret $\boldsymbol{\theta}$ as a cultural parameter: how much emphasis is placed on one's own interactions vs. the "consensus" of society?
  - How does manipulation change as we change $\boldsymbol{\theta}$, under network preservations?
- When $\boldsymbol{\theta} = \theta\mathbf{1}$ (i.e., homogenous society), there are thresholds $0 < \underline{\theta} < \overline{\theta} < 1$:
  - Sheep ($\theta < \underline{\theta}$): no manipulation as long as there is at least one Bayesian.
  - Narcissist ($\theta > \overline{\theta}$): no manipulation - each agent plays an isolated game
  - Reasonable People ($\underline{\theta} < \theta < \overline{\theta}$): the network is susceptible - subset of DeGroot agents get **Good** to build reputation, whereas another subset get **Bad**.
- However, when the experience weights in the network are heterogenous, then most sheepish agents also get manipulated.

Random Networks: Motivation

- Practitioners often fit real-world data to a random network model: Erdos-Renyi, configuration model, stochastic-block network, scale-free network, etc.
- Can we analyze whether certain random networks are impervious or susceptible to manipulation?
- Example: complete network is unrealistic - every agent talks to every other agent. But does Erdos-Renyi random network resemble the properties of the complete network because they are equivalent ex-ante?
- Goal: does the expected network say anything about realized network in terms of manipulation?

Random Networks: Setup

- Take as given the symmetric matrix of link probabilities:

$$\bar{\boldsymbol{\rho}}_n = \begin{pmatrix} p_{11}(n) & p_{12}(n) & \dots & p_{1n}(n) \\ p_{21}(n) & p_{22}(n) & \dots & p_{2n}(n) \\ \dots & \dots & \dots & \dots \\ p_{n1}(n) & p_{n2}(n) & \dots & p_{nn}(n) \end{pmatrix}$$

- Links are undirected: if i is linked to j, then j is linked to i as well.
- Links are formed independently of each other, given the probability in $\bar{\boldsymbol{\rho}}_n$.
- Assume that $\boldsymbol{\theta}^{(n)}$ is given as a function of n and uniformly bounded away from $\mathbf{1}$.
- The random network $\tilde{\mathbf{A}}_n$ is formed:

$$\tilde{\alpha}_{ij}^{(n)} = \begin{cases} (1-\theta_i^{(n)})/d_i^{(n)}, & \text{if there is a link } j \to i \\ 0, & \text{otherwise} \end{cases}$$

where $d_i^{(n)}$ is the degree of agent i.

Random Networks Theory: Conditions

- Want to analyze the "expected" network $\bar{\mathbf{A}}_n$ instead, given by $\bar{\alpha}_{ij}^{(n)} = (1-\theta_i^{(n)})p_{ij}(n)/\bar{d}_i^{(n)}$. What conditions must we impose?
    - Related to the conditions in Dasaratha (2019), but slightly more complex because: (i) stubborn Bayesian agents and (ii) normalized adjacency (i.e., "random walk") matrix.
- Non-vanishing spectral gap: The DeGroot-DeGroot submatrix of $\bar{\boldsymbol{\rho}}_n\bar{\mathbf{D}}_n^{-1}$ must have its second eigenvalue bounded away from its largest eigenvalue.
- Large expected degrees: All agents i have $\lim_{n\to\infty}\bar{d}_i^{(n)}/\log n = \infty$.
- Normal society: All θ's are grow/decay asymptotically at the same rate: for some ν, $\limsup_{n\to\infty}\theta_i^{(n)}/\theta_j^{(n)} \leqslant \nu$ for any DeGroots i, j.
    - Recall that in the heterogenous θ populations, the most sheepish agents get manipulated; in non-normal societies this applies too. The least trivial case is therefore when this assumption holds.

Random Networks Theory: Main Result

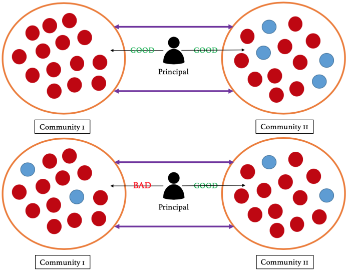
Theorem 3. (Random Networks)

Suppose that a sequence of $\bar{\boldsymbol{\rho}}_n$ has a non-vanishing spectral gap and satisfies the expected-degree condition; also let $\boldsymbol{\theta}$ be a normal society. For almost all ϵ, the random network $\tilde{\mathbf{A}}_n$ is impervious (resp. susceptible) if and only if $\bar{\mathbf{A}}_n$ is impervious (resp. susceptible) with high probability.

- Three key steps in the proof:
  - Step 1: Show that for any $\psi > 0$, $\lim_{n\to\infty} \mathbb{P}\left[||\tilde{\boldsymbol{\rho}}_n \tilde{\mathbf{D}}_n^{-1} - \bar{\boldsymbol{\rho}}_n \bar{\mathbf{D}}_n^{-1}||_2 \geqslant \psi\right] = 0$; that is, the norm of the difference between realized and expected RW matrices is small with high probability (expected-degrees condition).
  - Step 2: For any slacks γ_n, we have that $||\mathcal{D}(\gamma_n) - \mathbb{E}[\mathcal{D}(\gamma_n)]||_2$ is also small with high probability (normal society condition).
  - Step 3: Show that DeGroot network is connected with high probability, and so the set of optimal γ is the same under the realized and expected networks with high probability: $\tilde{\boldsymbol{\Gamma}}^* = \bar{\boldsymbol{\Gamma}}^*$ (non-vanishing spectral gap).

Application: Weak Homophily with Two Islands

- Weak homophily model: k groups with proportion of the population n: $(s_1, \ldots, s_k)$. Within-group link probability is p_s and between-group link probability is p_d.

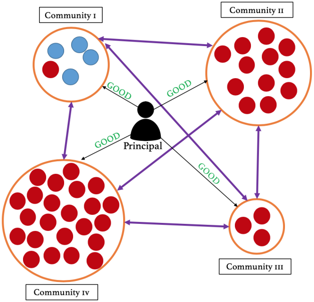


- With two communities of the same size $(s_1 = s_2 = 1/2)$:
  - Decreasing p_s may lead to more manipulation, but increasing p_s never does.
  - Increasing p_d may lead to more manipulation, but decreasing p_d never does.
  - If there are m Bayesian agents, then $m_1 = m_2 = m/2$ is worst distribution.

Optimal Seeding

- Problem: If one can endow a limited number of agents with Bayesian abilities (e.g., by educating them), which do you select?
- *Ring Network*: Clustered Bayesians is the worst-case, need to sprinkle them throughout the ring to get imperviousness.
- *Weak Homophily with Two Islands*: Evenly distributed Bayesians across both islands does the worst - better to concentrate them on a single island.
- Tradeoff between minimizing the DeGroot diameter of the network (like the ring) and having many DeGroot agents with beliefs close to the truth (like the weak homophily).

Application: Weak Homophily Imperviousness

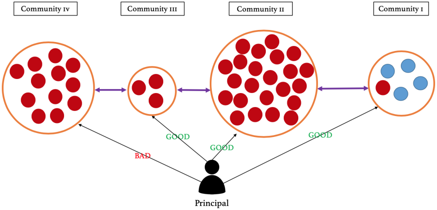


Proposition 7. (Weak Homophily)

Fix (p_s, p_d). There exists $\beta > 1$ such that as $n \to \infty$, if there are $m = O(\epsilon^{-1}(\beta - p_s + p_d)^{-1})$ Bayesian agents anywhere, then any weak homophily network with communities $\{s_\ell\}_{\ell=1}^{k}$ is impervious with high probability. In particular, if the number of Bayesian agents grows unboundedly with n, any weak homophily network is impervious for any ϵ, with high probability.

Application: Strong Homophily

- Strong homophily: Each community has a quality score $\lambda_j \in \mathbb{R}$. Within-group link probability is still p_s, but between-group link probability is p_d only for the two nearest communities (and otherwise 0).



Proposition 8. (Strong Homophily)

Fix (p_s, p_d). There exists $\epsilon > 0$ such that for every large n, there is a strongly assortative homophily network $\tilde{\mathbf{A}}_n$ with communities $\{s_\ell\}_{\ell=1}^{k}$ susceptible to manipulation with high probability. On the other hand, the weak homophily network with the same commnunities $\{s_\ell\}_{\ell=1}^{k}$ is impervious with high probability.

Conclusion

- Embed the classical reputation setup in a social learning environment where agents communicate with either each other as they interact with the principal.
- The social network is heterogenous on multiple dimensions: reasoning sophistication, network position, personal experience interpretation, etc.
- Principal can sometimes exploit his reputation to increase payoffs, but depends on network structure, the learning mechanisms employed by the agents, and societal norms.
  - Focus of DeGroot learning is how badly it can perform (in worst-case) relative to information aggregation benchmark.
  - Principal is not an adversary: understand how bad learning is when the principal strategically chooses underlying state.
- Main prescription: focus should be on increasing communication between groups who do not normally communicate; this is more critical than improving the sophistication of many agents.

Future Work

- What happens when Bayesian agents are also strategic in their communications?
  - For example, add an action for the principal **Super Good** which is costly but gives a high payoff to the agent. Principal plays this conditional on the Bayesian spreading positive beliefs about the reputation of the principal (instead of truth).
- Allow the principal to play mixed strategies and/or time-varying strategies: can only lead to more manipulation.
  - Main challenge in the former is that DeGroot beliefs follow a weird limiting distribution and beliefs are network correlated based on recent realizations of the mixed strategy.
  - Conjecture the latter reduces to time-invariance when allowed to play correlated mixed strategies (i.e., convex hulls are the same), assuming the principal does not observe the realizations of his mixed strategy.
- Experimental considerations: testable hypotheses about principal-agent(s) interactions in a social setting.