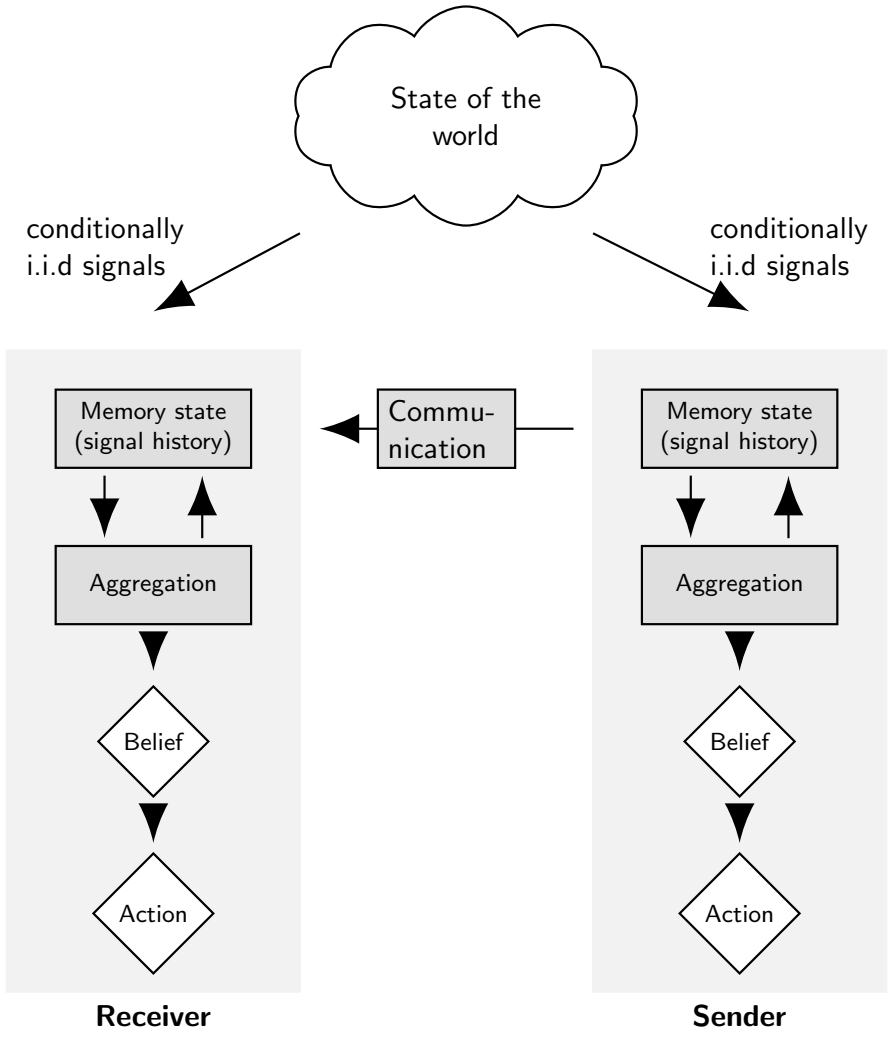
Social Learning

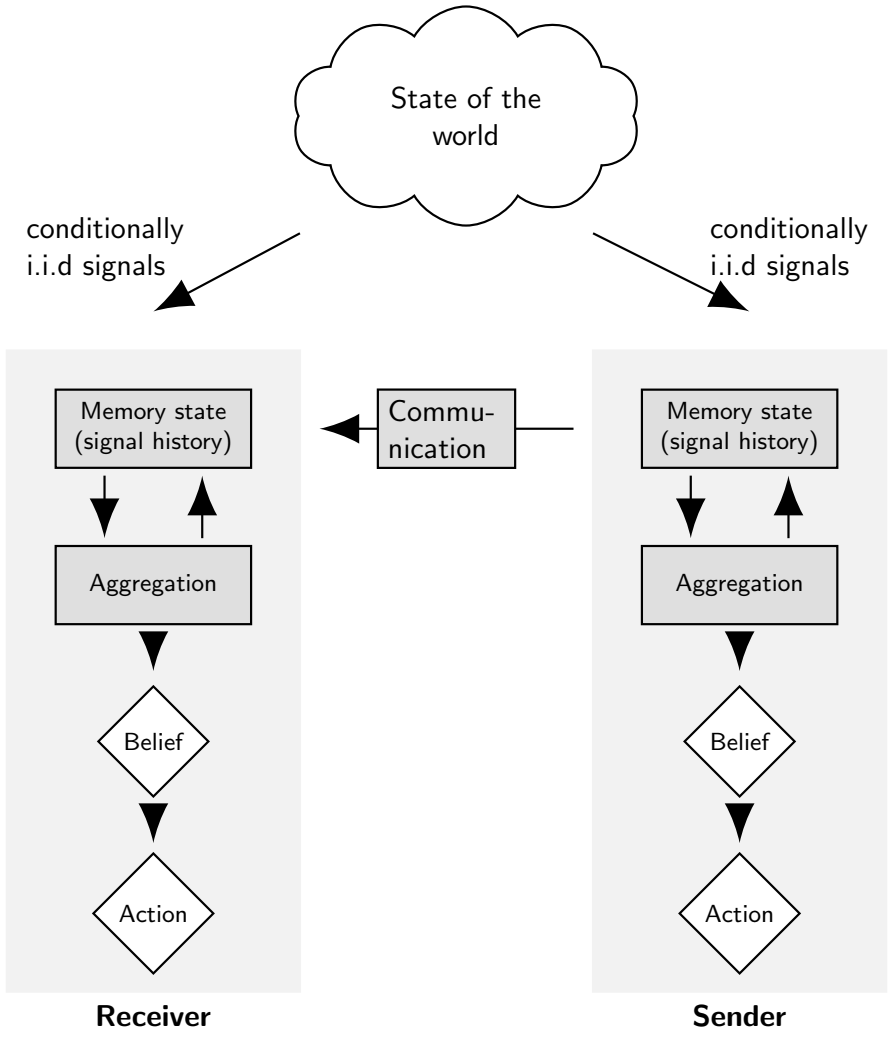
Markus Mobius
Microsoft Research

March 28, 2014

Social Learning Framework



Social Learning Framework



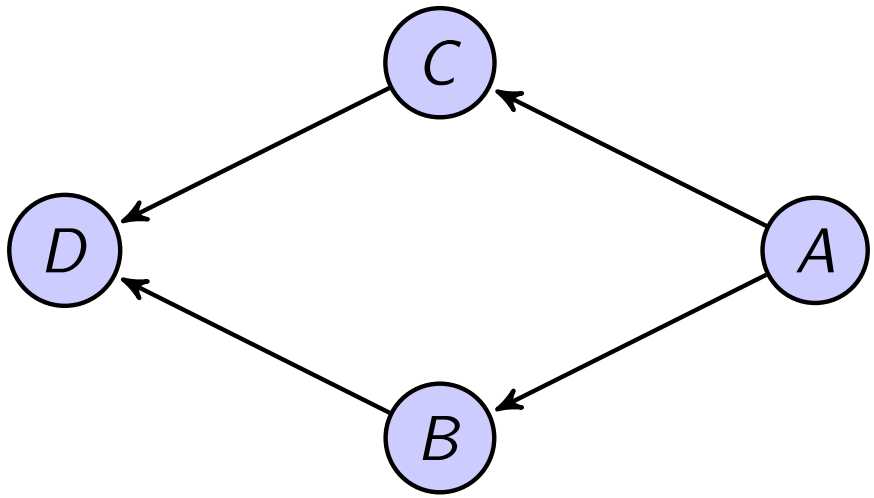
Elements

- directed or symmetric social network of n agents
- $(i, j) \in g$ means that there is a link between agents i and j
- some true state of the world θ (binary or continuous)
- every agent in the network is assumed to have some signal x_i about the state of the world (also binary or continuous)
- Each agent's objective is form beliefs about the true state of the world that are as accurate as possible.

Conversations

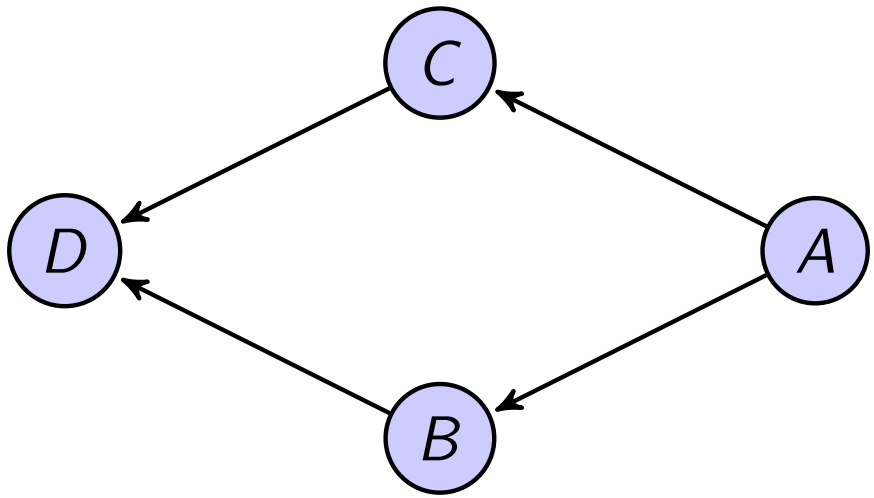
- Time is usually discrete and agents can talk to a subset of their neighbors in each period.
- There is a *conversation network* $g_t^C \subset g$ such that $(i,j) \in g_t^C$ implies that i sends information to j (hence, j listens to i).
- Many theoretical and most empirical papers do not distinguish between the social network and the conversation network such that $g_t^C = g$.

Challenges



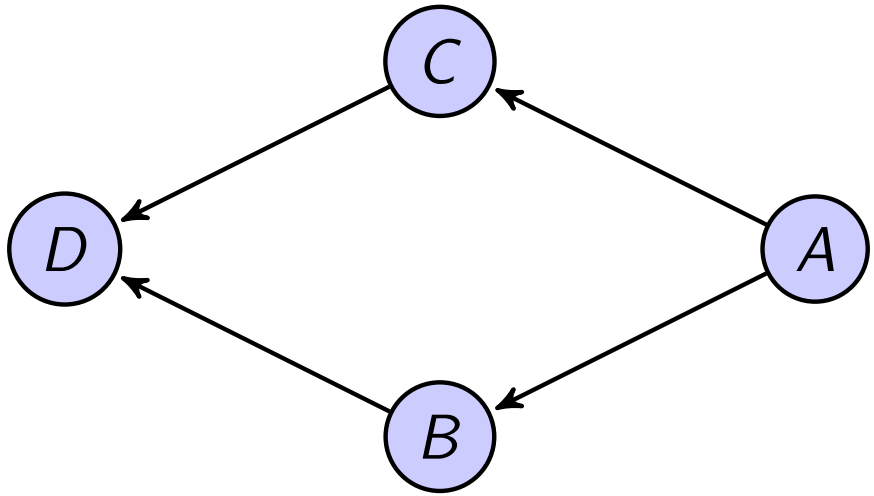
- What do agents talk about?
- What do they store in memory?
- How do they process messages?

"Streams" model (efficient)



- In period $t = 1$, agent B tells neighbor D: "my signal is x_B".
- In period $t = 2$, agent B tells D: "I heard that A has signal x_A".

"Streams" model: disadvantages



- memory intensive, complex
- most models assume that agents transmit *summary statistics* in each time period
  - actions
  - posterior beliefs
- summary statistics simplify messaging but increase computational burden of rational agents
  - lack of independence

Modeling approaches

1. Bayesian models
   - typically complete social learning in long run
   - experimental evidence suggests that real agents use simpler heuristics
2. Naive learning (DeGroot)
   - surprisingly good aggregation
3. Diffusion models
   - appropriate when most agents have no information
   - percolation rather than signal aggregation

Observational Learning

Bikhchandani, Sushil, Hirshleifer, and Welch (1992), Banerjee (1992)

- agents observe their neighbors actions
- actions are *coarse* (typical there are just 2 actions)
- line network
- Agent t can observe the actions of all agents before her.
- Directed conversation network: agent t can observe the action of agent $t-1$ but not vice versa (important!)
- state of the world is either H or L and both states are a priori equally likely
- Each agent observes a conditionally i.i.d. binary signal $x_i \in \{H, L\}$ which is correct with probability p.
- Agents can choose a binary action $a_i \in \{H, L\}$ in every period and they get utility 1 from choosing an action that matches the state and they get utility 0 otherwise.

Observational Learning: herding

- actions will converge almost surely as $t \to \infty$ but beliefs will not converge
- actions can converge with positive probability to the *wrong* action.
- Consider the first agent's decision problem: WLOG assume that she has an H signal and her action will therefore imitate her signal.
- Now consider second agent: she can infer the first agent's signal from her action.
  - If she has a high signal herself then she will take action H.
  - Otherwise, her signal just cancels the signal of her predecessor in which case she will randomize.
- Finally, consider the third agent. If she sees both of her predecessors take action H then she will imitate that action *regardless* of her signal.

Observational Learning: necessary conditions for herding

- "chunky" actions
- bounded signals (Smith and Sorenson, 2000)
- directed conversation network (Acemoglu, Dahleh, Lobel,Ozdaglar, 2011; Mossel and Tamuz 2013)

Learning from Posteriors

DeMarzo, Vayanos and Zwiebel (2003)

- Bayesian learning in a strongly connected social network
- conversations take place along all edges in every time period ($g_t^C = g$)
- agents communicate Bayesian posteriors
- result: all agents learn everyone's signal after at most n^2 periods
  - this upper bound only depends on connectedness and the size of the network but not on the structure of the social graph

Experimental Evidence for Bayesian Social Learning

- Bayesian models of social learning generally assume that agents know the full conversation network in order to draw proper inferences.
- This is clearly a strong assumption given that agents in a social network typically have poor knowledge of even second-order neighbors.
- People do not always process information as perfect Bayesians even when they process information in isolation.
  - cognitive psychologists and economists have shown this through simple laboratory experiments such as ball-and-urn problems

Experimental Evidence for Bayesian Social Learning

- Weizsacker (2010) shows in a meta-study of 13 observational learning experiments that agents hesitate to follow others and ignore their own signal unless the strength of their neighbors' information is very strong (with a likelihood ratio of at least $2:1$).
- This suggests that agents incorporate *too little* information compared to the Bayesian benchmark.
- Interestingly, Kariv (2004) observe herding in their laboratory experiments but rarely on the incorrect action.

Naive Bayesian

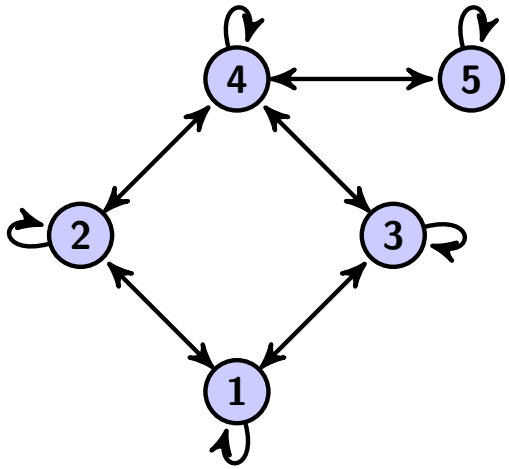
- The naive learning literature replaces Bayes rule with a simple Markovian heuristic rule.
- workhorse: DeGroot model with continuous signals x_i where agents form a guess x_i^t about the state of the world in every period as follows:
  - their best guess before talking to anyone is simply their signal such that $x_i^0 = x_i$.
  - In every subsequent period, they average their own last guess with the guesses of the neighbors they have listened to, such that:

$$x_i^t = \frac{x_i^{t-1} + \sum_{(j,i) \in g_t^C} x_j^{t-1}}{1 + \left| \{(j,i) | (j,i) \in g_t^C \} \right|} \tag{1}$$

Naive Bayesian: justification

- This heuristic averaging rule can be justified as a form of "naive" Bayesian learning as follows.
- Assume that the state of the world is drawn from a normal distribution with precision h_θ.
- For simplicity, we assume throughout that precision is so low that we can ignore it for calculating posteriors ($h_\theta \approx 0$).
- Every agent observes a signal $x_i = \theta + \epsilon_i$ where ϵ_i is a normally distributed error term with mean zero and precision h_ϵ.
- The DeGroot rule is then the correct Bayesian updating formula in period $t = 1$.
- However, in subsequent periods it "double-counts" neighbors signals as discussed above.

DeGroot: example



Denote the vector of guesses at time t with $x^t = (x_i^t)$. It is easy to see that $x^{t+1} = Mx^t$ where:

$$M = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{2}$$

DeGroot: convergence of beliefs

- matrix M is right-stochastic because its rows sum up to 1
- matrix is irreducible because the associated graph is strongly connected
- matrix is aperiodic because every agent listens to herself
- Perron-Frobenius: largest eigenvalue of the matrix is 1 and there is a unique left eigenvector π with positive components such that:

$$\pi M = \pi \tag{3}$$

DeGroot: convergence of beliefs

We next show that the guesses of all agents converge to $x_i^\infty = \sum_i \pi_i x_i$.

- We can write $x^t = M^t x^0$.
- Assume that we want to focus on the guess of the first agent at time t. We can write $x_1^t = e_1 x^t$ where $e_1 = (1, 0, .., 0)$. This implies:

$$x_1^t = e_1 M^t x^0 \tag{4}$$

- We know that $e_1 M^t \to \pi$ because the Markov chain associated with M is ergodic. Hence, we obtain:

$$x_1^t \to \pi x^0 \quad \text{as } t \to \infty \tag{5}$$

- By applying the same trick to every agent's belief we can show that all beliefs converge to $\sum_i \pi_i x_i$.

DeGroot: social influence

The vector π captures the *social influence* of every agent: the opinions of agents with greater influence has a greater weight in the final converged belief.

- In the case of symmetric networks it is easy to show directly that social influence is proportional to the agent's degree, d_i (plus 1):

$$\pi_i \sim 1 + d_i \tag{6}$$

- We can immediately deduce that social learning in the DeGroot model is not efficient in the sense that the best possible guess with normally distributed signals is $\frac{1}{n} x_i$.

DeGroot: social influence

The vector π captures the *social influence* of every agent: the opinions of agents with greater influence has a greater weight in the final converged belief.

- However, converged beliefs in the DeGroot model will typically be extremely close to the efficient guess provided that all agents in the network have "small degree" compared to the size of the network n.
- One sufficient condition for "smallness" is that there is a maximum degree d_{max} regardless of the size of the network.
- In this case, a version of the law of large numbers holds and beliefs will converge in probability to the efficient outcome.
- Jackson and Golub (2010) provide a general analysis and hence a theoretical underpinning for the "wisdom of crowds" as popularized in the recent book by Surowiecki (2005).

DeGroot: speed of convergence

- An interesting feature of equation 6 is that social influence in a symmetric network is *only* determined by the degree distribution but not by any structural properties of the network such as cliquishness or average social distance.
- However, these features do affect the speed of convergence as Jackson and Golub (2012) demonstrate.

DeGroot: speed of convergence

- An easy way to see this is to think of the social network as a partition of disjoint social islands. Every agent has d_1 random intra-islands links and d_2 random inter-island links.
- If the share of intra-island links is proportional to the population share of the island then we have a simple random network. However, if the share of intra-island links is larger than the population share then the social network exhibits *homophily* because agents prefer friendships with own-island neighbors.
- It is easy to see that homophily does not alter agents' social influence and therefore has no effect on long-term learning. However, it can significantly slow down the speed of convergence: opinions will always quickly converge within an island, but it can take much longer to average inter-island differences in opinions.

Diffusion models

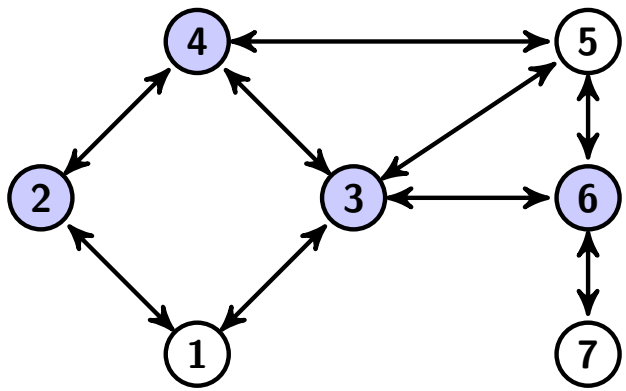
- Diffusion models are different from Bayesian and naive social learning models because there usually are no conflicting signals.
- Instead, we are interested in studying the percolation of information to uninformed agents.
- Diffusion models are usually very simple which makes them suitable for empirical applications where the parameters of the model can be estimated by using a simulated method of moments approach (Topa 2000, Banerjee et al. 2013).

Diffusion models: Jackson-Calvo Armengol (2004)

- We consider a continuous time Markov process on a symmetric social network g.
- At time t, the state of the system is described by a map $\eta^t : A \to \{0, 1\}$ where A is the set of agents.
- We say that agent i is employed at time t if $\eta^t(i) = 1$ and that she is unemployed otherwise.

Diffusion models: Jackson-Calvo Armengol (2004)

Example:



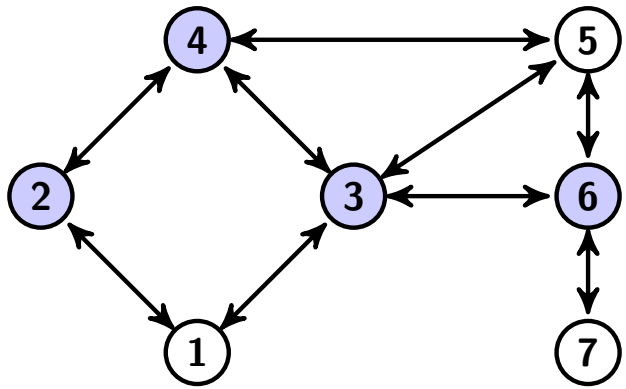
Note: Employed agents are darkly shaded.

Diffusion models: Jackson-Calvo Armengol (2004)

- All agents get job opportunities at rate a.
- If an unemployed agent gets an opportunity she will use it for herself.
- However, if an employed agent receives an opportunity she will tell one of her unemployed neighbors about it.
  - If several of them are unemployed, she will randomly choose one of them.
  - In many diffusion models state 1 (being informed/having a job etc.) is an absorbing state. In this model, agents become unemployed at an exogenous rate u.

Diffusion models: Jackson-Calvo Armengol (2004)

Example:



- agent 1 switches from 0 to 1 at rate $2.5a$
- agent 5 switches at rate $3a$
- agent 7 at rate $1.5a$

Being friends with employed agents increases the chances to find a job.

Diffusion models: attractive spin systems

- We now have a fully specified continuous-time Markov process on a finite state space (namely, the set of all configurations which has size 2^n).
- The process is therefore ergodic.
- Even though we cannot derive a closed-form solution for the ergodic distribution, we can exploit that it is an *attractive spin process* (Liggett 1985).
  - The switching rate from 0 to 1 for any agent is (weakly) increasing if any neighbor's state is switched from 0 to 1.
  - The switching rate from 1 to 0 is (weakly) decreasing if any neighbor's state is switched from 0 to 1.
- We now know that the states of any 2 agents are positively correlated.
- Moreover, there is a hysteresis effect: if we compare two otherwise equal communities, one with full employment and the other one with partial unemployment, then the latter community's unemployment rate will always be higher in expectation at any time $t > 0$.