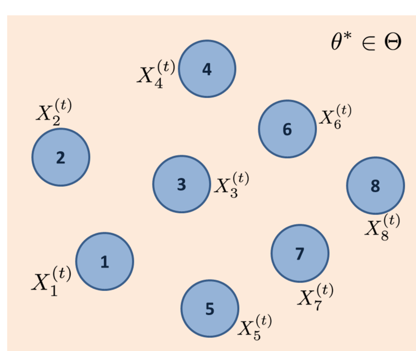
Social Learning and Distributed Hypothesis Testing

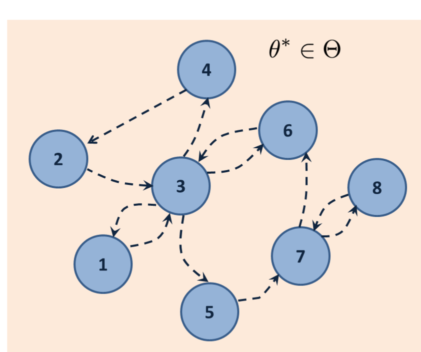
Anusha Lalitha[1], Tara Javidi[1], Anand Sarwate[2]
[1]Dept. of Electrical and Computer Engineering, University of California, San Diego
[2]Dept. of Electrical and Computer Engineering, Rutgers University

Model

- A set of "n" nodes and a finite set of "M" hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Nodes are make observations $X_i^{(t)}$ at every time instant.
- Observations are statistically governed by fixed true hypothesis θ*.



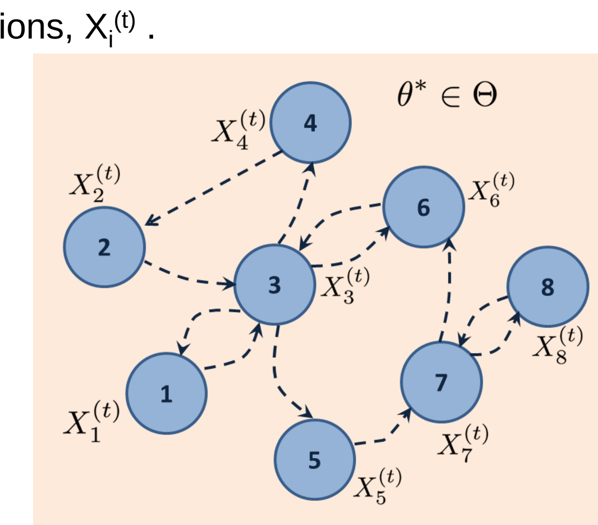
- Each node knows the conditional likelihood functions under each hypothesis. The set is given as $\{f_i(.;\theta): \theta \in \Theta\}$
- Nodes are connected in a strongly connected network.



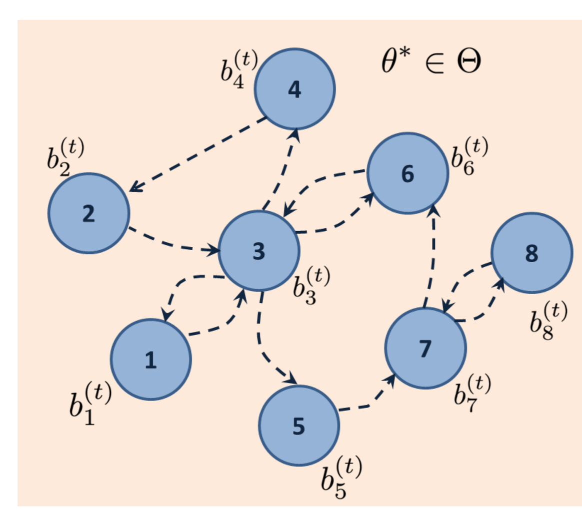
Goal is the parametric inference of the fixed unknown global hypothesis θ* with *n* nodes.

Learning Rule

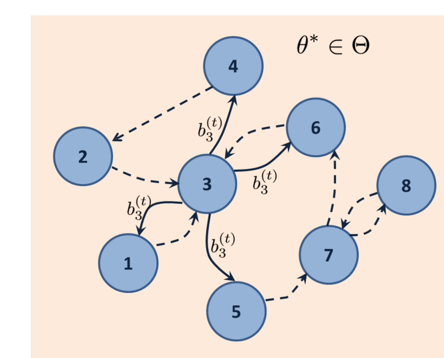
- Every node begins with a strictly positive prior on each hypothesis.
- **Making observation:** At every time instant (t>0), nodes are make observations, $X_i^{(t)}$.

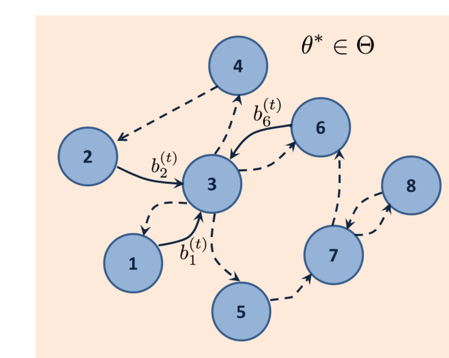


- **Bayesian update of belief:** Every node using its observations performs a Bayesian update of its belief vector, $b_i^{(t)}$ (.), which is probability vector on the set of hypotheses.

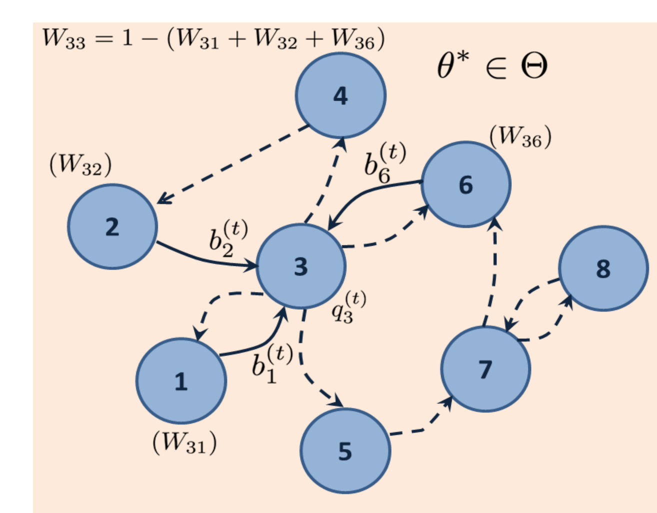


- **Exchanging beliefs through network:** Each node then sends its belief vector to the neighboring nodes. Similarly receives the belief vectors of its neighbours.





- **Updating estimate as average of log-beliefs:** Using the belief vectors from the neighbors, each node updates its local estimate vector , $q_i^{(t)}$ (.), which is also a probability vector on the set of hypotheses. Estimate vector is computed as the average of log-beliefs using weights.



Rate of Convergence

- Weight $W_{ij} > 0$ if and only if there is an edge from node i to node j and $W_{ii} = 1 - \sum_{j=1}^{n} W_{ij}$.
- Markov chain defined by weight matrix W is irreducible.
- The Markov chain defined by W has a unique stationary distribution, $\mathbf{v} = \{v_1, v_2, .., v_n\}$, which is the left eigenvector of W associated with eigenvalue 1.
- For any θ not equal to true hypothesis, we define the network divergence as

$$K(\theta^*, \theta) = \sum_{j=1}^{n} v_j D(f_j(.;\theta^*) || f_j(.;\theta))$$

Theorem:
For every node in the network, the estimate of wrong hypothesis θ ≠ θ* computed using the proposed learning rule converges to zero exponentially fast with probability one. Furthermore, the rate of rejecting hypothesis θ in favor of θ* is given by the network divergence between θ and θ*.

$$\lim_{t\to\infty} \frac{1}{t} \log q_i^{(t)}(\theta) = K(\theta^*, \theta)$$

Factors influencing convergence

Rate of rejecting θ in favor of θ* is influenced by eigenvector centrality and the KL-divergence between conditional likelihood functions of θ and θ*.

- Set $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and $\theta^* = \theta_1$
- If node i is connected to node j

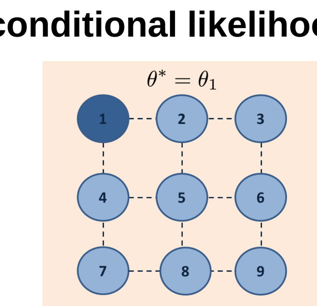
$$W_{ij} = \frac{1}{\text{degree of node } i}$$

- Otherwise 0.
- Node i is uninformed if

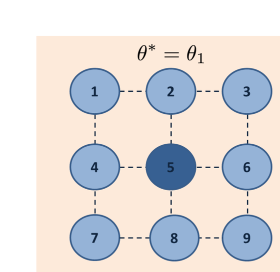
$$f_i(\cdot;\theta^*) = f_i(\cdot;\theta) \text{ for some } \theta \neq \theta^*$$

- Node i is informed if

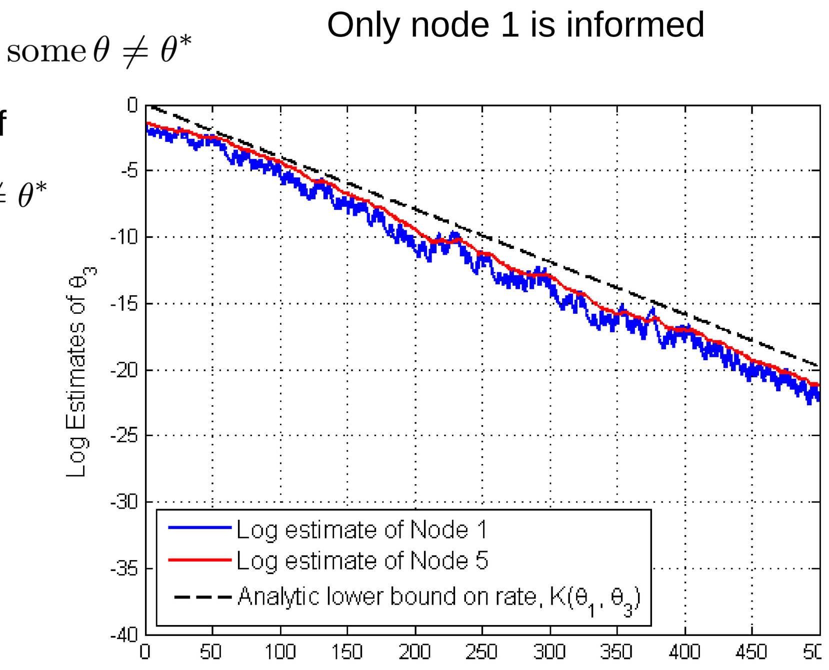
$$f_i(\cdot;\theta^*) \neq f_i(\cdot;\theta) \, \forall \theta \neq \theta^*$$

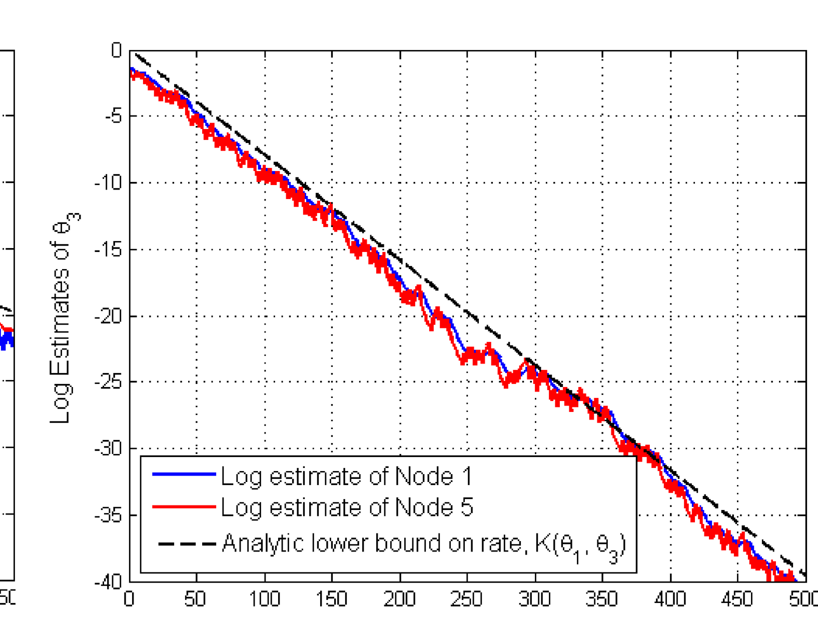


Only node 1 is informed



Only node 5 is informed



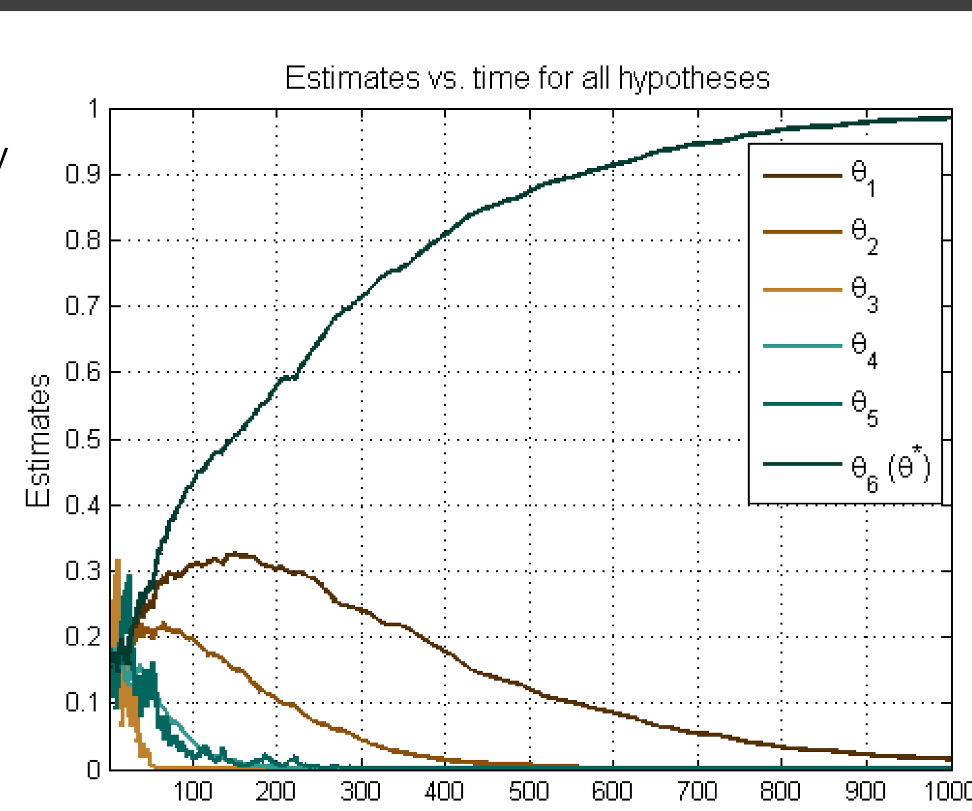


Learning in a Co-authorship network

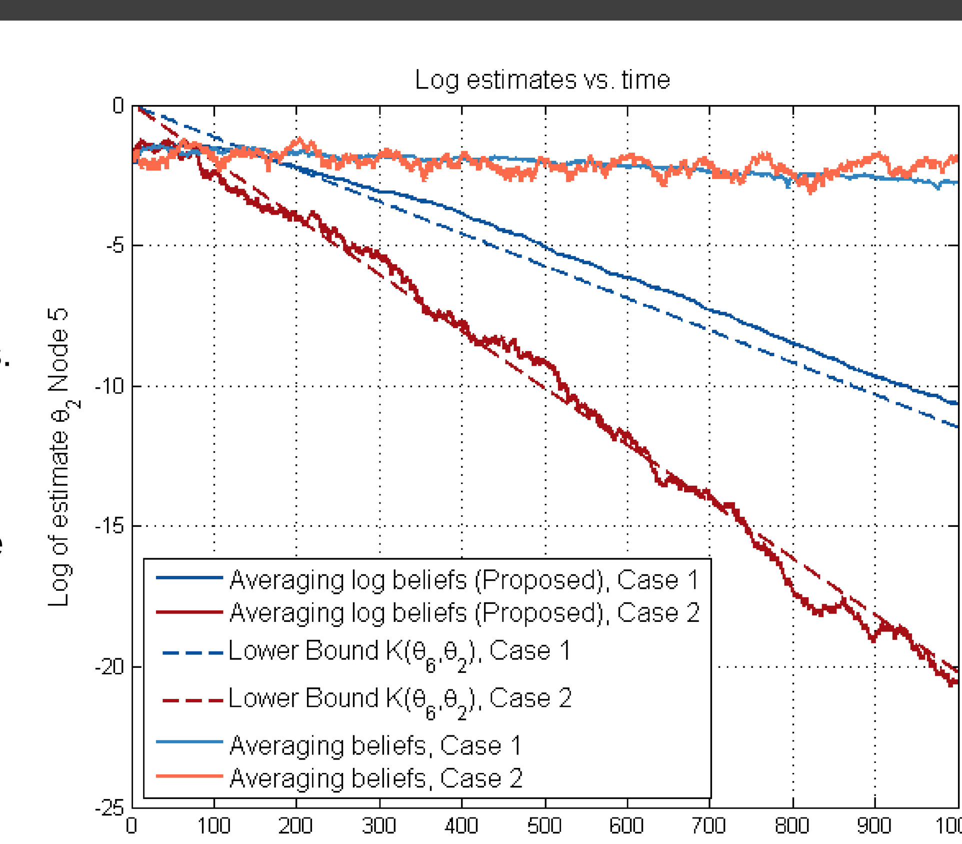
- Network is an undirected weighted network of co-authorships formed from high-energy theory collaborations over a period of 5 years.
- Weights capture strength of collaboration [3].
- Set of hypotheses is

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_6\} \text{ and } \theta^* = \theta_6$$

- Case 1:
  - Scientists are randomly divided into 5 groups.
  - Each group can distinguish between θ_6 and some $\theta \neq \theta_6$
- None of them can individually learn θ_6 but using the proposed rule, nodes learn θ_6, as shown in adjacent figure.



- We compare our rule with existing one in the literature where averaging of log-beliefs is replaced with averaging the beliefs [2].
- We compare the performance in two cases.
- Case 2:
  - Scientists are randomly divided into 10 groups.
  - Each group can distinguish between θ_6 and some $\theta_i \neq \theta_6$ and some $\theta_j \neq \theta_6$
- Our rule performs faster than the previous rule in both the cases.
- We show this theoretically in [1].



References:
1. A. Jadbababie, P.Molavi, and A. Tahbaz-Salehi, "Information Heterogeneity and the speed of learning in social networks", Columbia Business School Research Paper No.13-28, June 2013.
2. A. Lalitha, A. Sarwate and T. Javidi, "Social learning and distributed hypothesis testing", in Proceedings of International Symposium on Information Theory (ISIT), July 2014.
3. M. E. J Newman, "Scientific collaboration networks ii. Shortest paths, weighted networks and centrality ", Phys. Rev. E, vol. 64, p. 016132, June 2001.