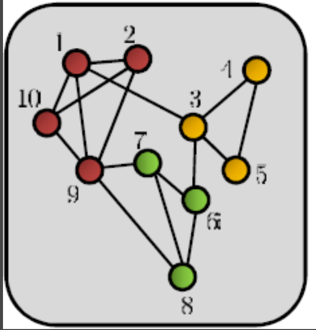
# Adaptation & Learning by Networked Agents

Ali H. Sayed
UCLA Electrical Engineering



# Survey Articles



CONTRIBUTED PAPER

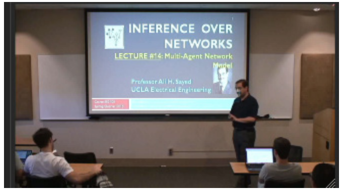
Sayed: Adaptive Networks

460 PROCEEDINGS OF THE IEEE | Vol. 102, No. 4, April 2014

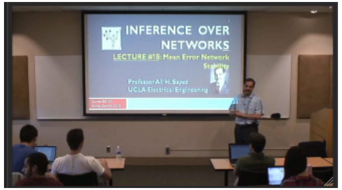
## Adaptive Networks

By ALI H. SAYED, *Fellow IEEE*

**Video Lectures** (*Inference over Networks* @ UCLA)








IEEE Signal Processing Magazine, May 2013

## Diffusion Strategies for Adaptation and Learning over Networks


## Adaptation, Learning, and Optimization over Networks

**Ali H. Sayed**
University of California at Los Angeles

Foundations and Trends® in Machine Learning
Volume 7, Issue 4-5, 2014

now
the essence of knowledge
Boston — Delft

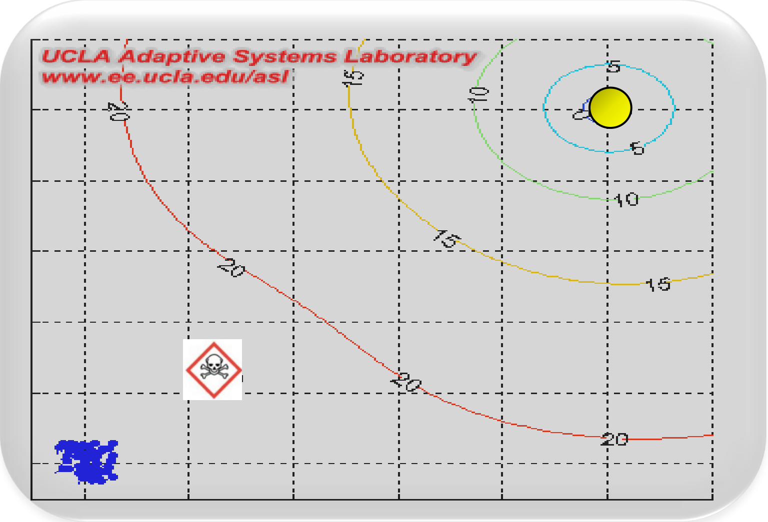
# Distributed Inference & Processing



**Deals with the discovery of global information from local interactions among dispersed agents.**

**Features:**

- Coordination;
- In-network processing;
- Dispersed agents.

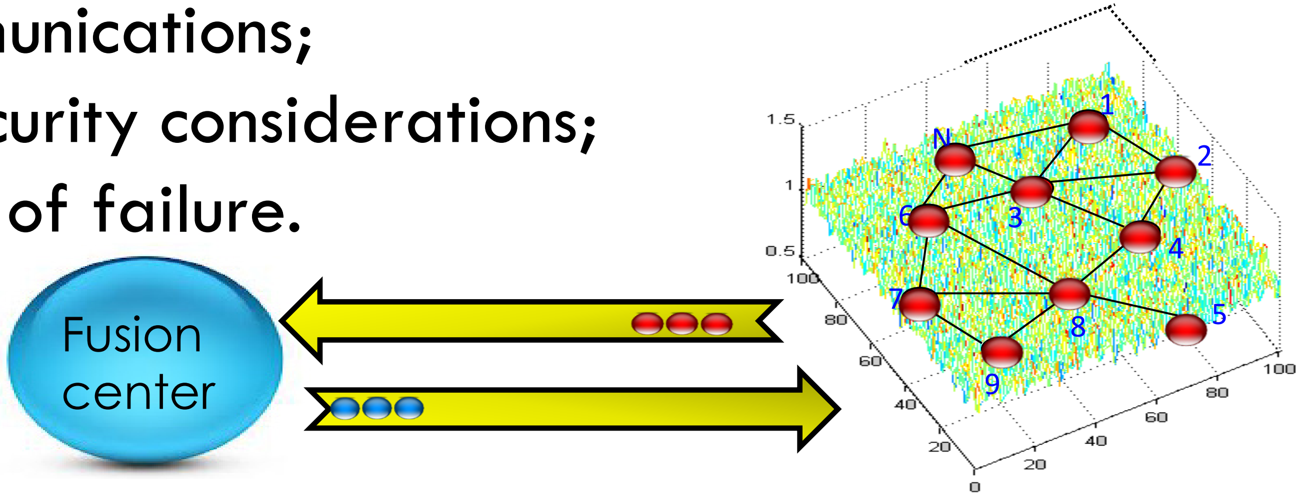


# Centralized Processing



**Exchange of data between the dispersed agents and a fusion center.**

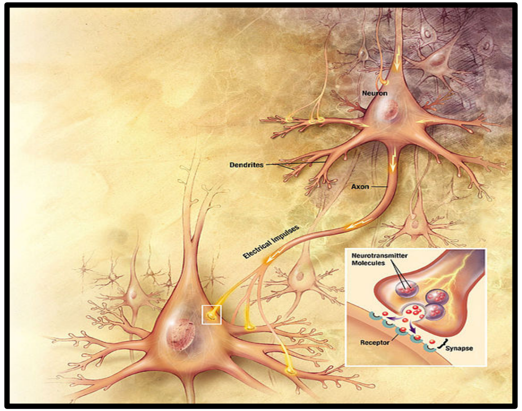
- Cost of communications;
- Privacy & security considerations;
- Critical point of failure.



# Biological Networks



## Nature provides splendid examples of real-time decentralized learning & adaptation.




**Source:** Wikimedia.

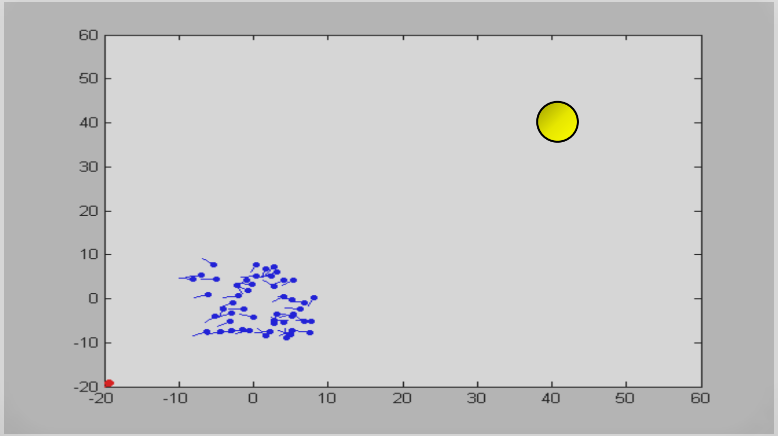
**Source:** Wikimedia; Creative Commons License.

**Source:** S. Pratt Lab, ASU.

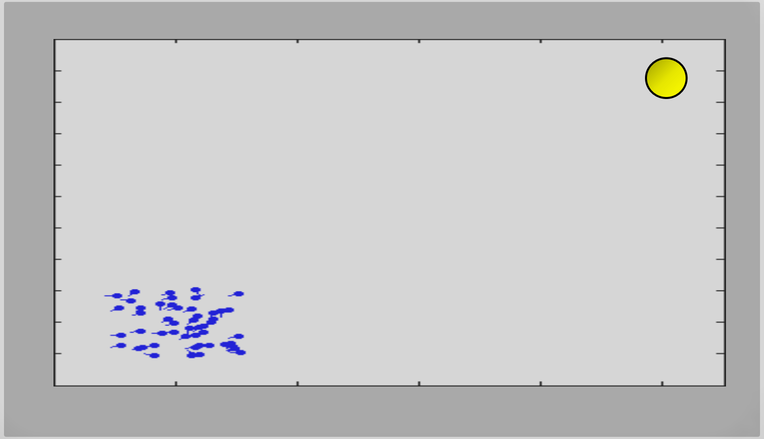
# Bio-Inspired Networked Cognition



**network cooperation**



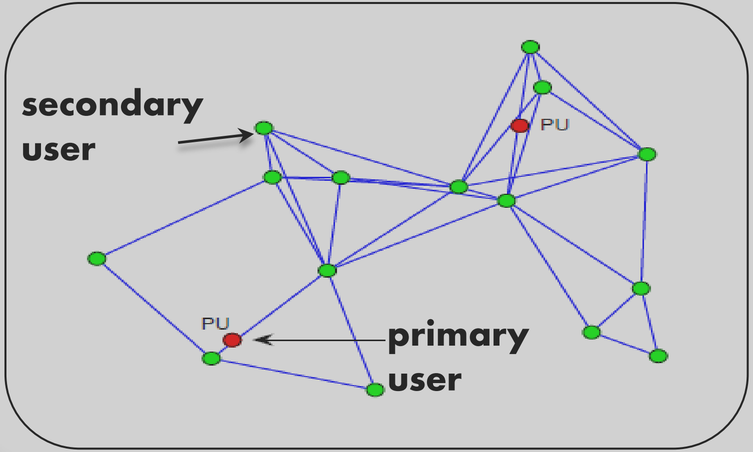

**network competition**

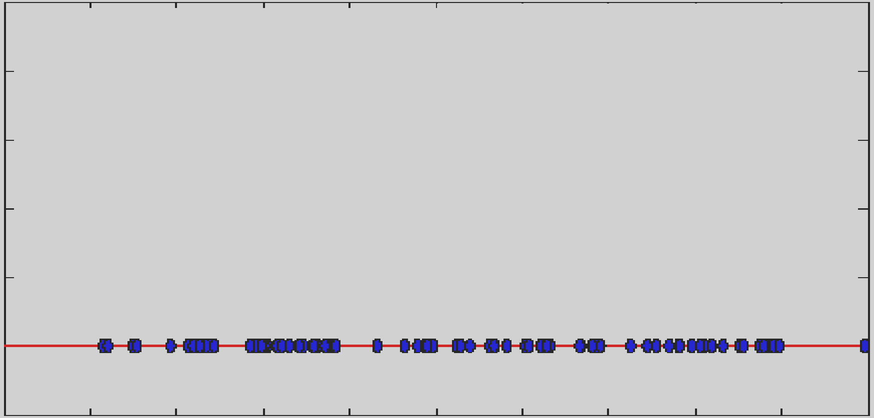


**Sayed** (*Proc. IEEE*, April 2014)
**Tu & Sayed** (*IEEE Trans. Sig. Process.*, August 2011)

# Bio-Inspired Radio Mechanism





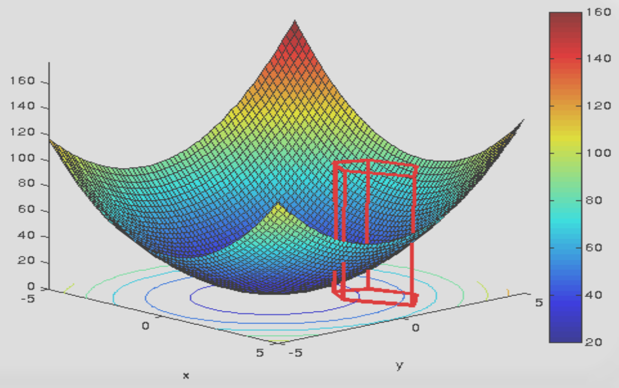
**Available frequency bands** → modeled as food sources.

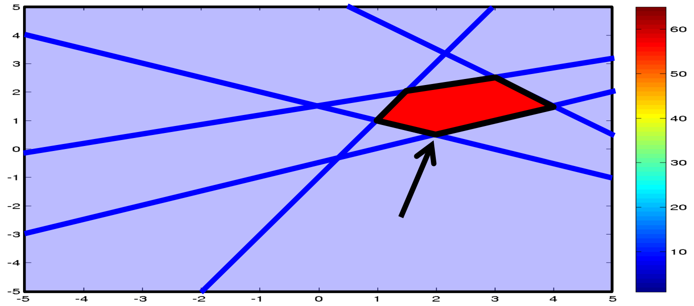
**Occupied PU bands** → modeled as obstacles.

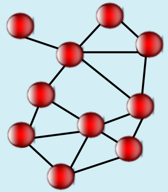
**Lorenzo, Barbarossa & Sayed** (*IEEE Trans. Sig. Process.*, June 2013)

# Optimization & Tracking



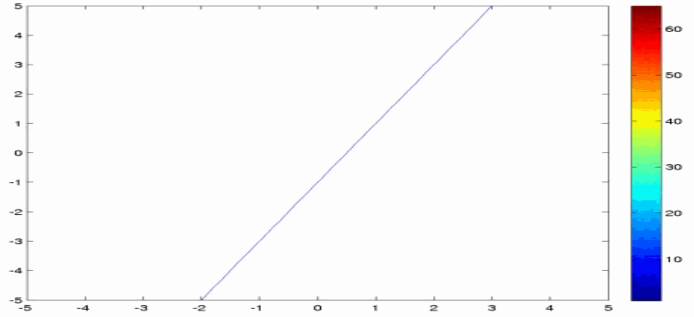






$$\min_{w} \sum_{k=1}^{N} J_k(w)$$

$$\text{subject to} \begin{cases} g_k(w) \le 0 \\ h_k(w) = 0 \end{cases}$$



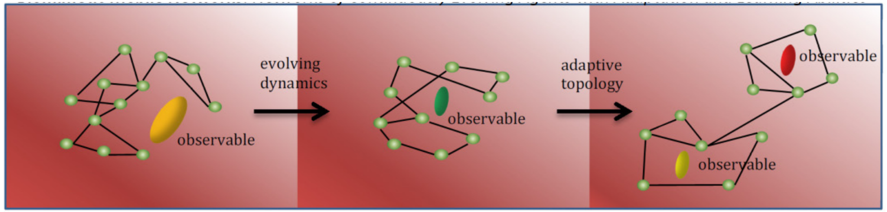
**Towfic & Sayed** (*IEEE Trans. Sig. Process.*, August 2014)

# Why Networked Solutions?



- Data already available at dispersed locations (cloud).
- Power of cooperation → mining of Big Data sets.
- Privacy and security considerations.
- Robustness and resilience (biological networks).
- Robotic swarms (disaster areas).
- **Apps:** Social networks; healthcare informatics; smart manufacturing, mobile health, transportation, energy,…

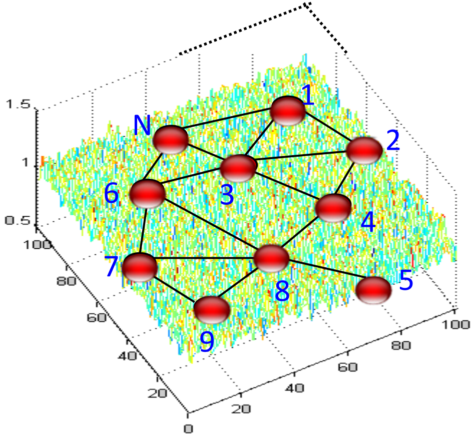
**Exploratory mobile agents or swarms**



# Adaptive Networks: Opportunities



- **Adaptive agents:** learn from streaming data.
- **Cooperative agents:** interact locally.
- **Adaptive topology:** re-wire the graph.
- **Distributed learning and inference.**




**New Degrees of Freedom:**

→ Coordination

→ Topology

→ Mobility

# Adaptive Networks: *Challenges*



**Challenges:**

→ Coupled dynamics

→ Selfish behavior

→ Privacy & secrecy

→ Asymmetries

**"Too many cooks spoil the broth."**

**"All animals (agents) are equal, but some animals (agents) are more equal than others."**

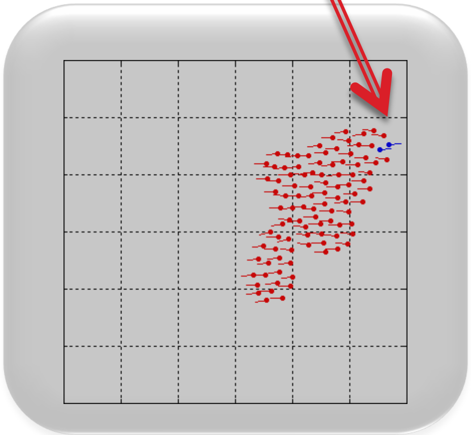
# Not all agents are equal…



## Quorum response in animal groups

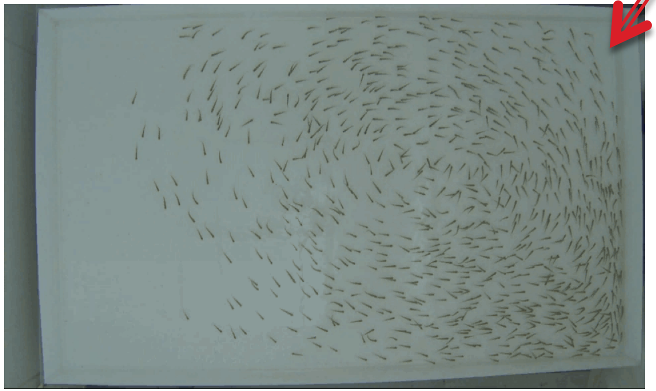
(**Sumpter & Pratt**, *Phil. Trans. R. Soc.*, 2009)

Observe sudden turn
at this corner



**Tu & Sayed**, *Proc. Cog. Inform. Process.*, (2012)

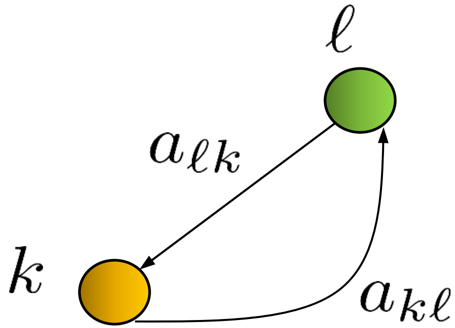
Observe sudden turn
at this corner



Animal Behavior Lab (Couzin, Princeton University)

# Interactions: *State-Dependent*






$$a_{\ell k} \propto \text{Prob}\left(\mathbb{I}_\ell = 1 | \text{state of } \ell\right)$$

**Probability of being an informed agent.**

# Regularization; Sparsity; Priors, ..



$$w^o \overset{\Delta}{=} \arg\min_w \; \underbrace{R(w)}_{\textbf{regularization}} + J(w)$$

**Examples:**

$$R(w) = \begin{cases} \rho\|w\|^2, & (\ell_2\text{–regularization}) \\ \alpha\|w\|_1, & (\ell_1\text{–regularization}) \\ \alpha\|w\|_1 + \rho\|w\|^2, & (\text{elastic-net regularization}) \end{cases}$$

# Risk and Loss Functions



$$w^o \triangleq \arg\min_{w} \; R(w) + \underbrace{J(w)}_{\textbf{risk}}$$

**Often:**

$$J(w) = \mathbb{E}\,\underbrace{Q(\boldsymbol{h}^{\mathsf{T}} w, \boldsymbol{\gamma})}_{\textbf{loss}}$$

**Data:**
(labels, features)
(reference, regressors)

# MSE, SVM, Perceptron, Boosting, …



**Examples:**

$$
Q(\boldsymbol{h}^{\mathsf{T}}w,\gamma) = \begin{cases}
(\gamma - \boldsymbol{h}^{\mathsf{T}}w)^2, & \text{(quadratic loss)} \\
\ln\left(1 + e^{-\gamma \boldsymbol{h}^{\mathsf{T}}w}\right), & \text{(logistic regression)} \\
e^{-\gamma \boldsymbol{h}^{\mathsf{T}}w}, & \text{(exponential loss)} \\
\max\{0,\, 1 - \gamma \boldsymbol{h}^{\mathsf{T}}w\}, & \text{(hinge loss)} \\
\max\{0,\, -\gamma \boldsymbol{h}^{\mathsf{T}}w\}, & \text{(Perceptron loss)}
\end{cases}
$$

# Classical Gradient-Descent

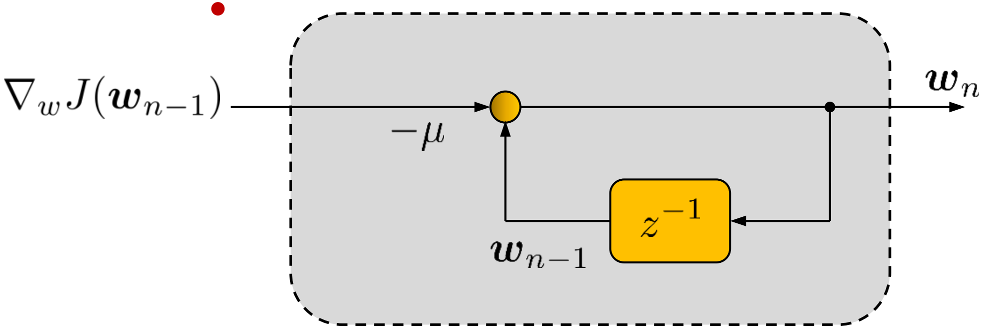


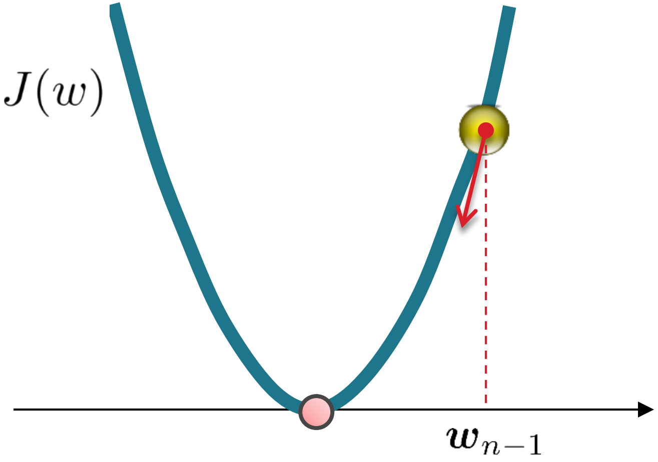
$$w^o \overset{\Delta}{=} \arg\min_{w} \; J(w)$$

$$J(w) = \mathbb{E}\, Q(\boldsymbol{h}^{\mathsf{T}} w, \boldsymbol{\gamma})$$

?

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} - \mu \nabla_w J(\boldsymbol{w}_{n-1})$$



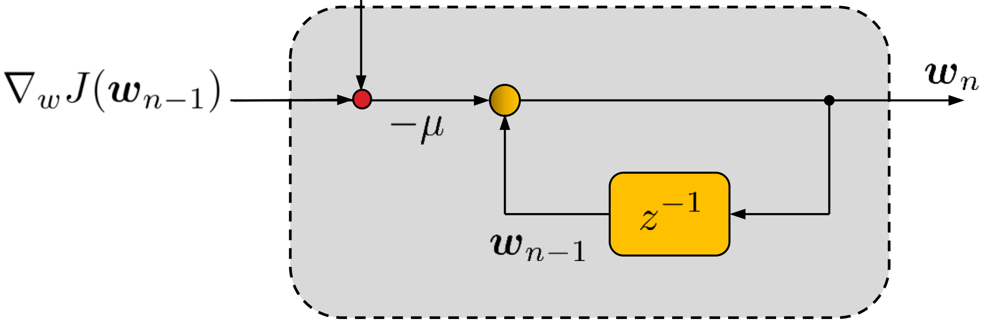
# Stochastic Gradient Learning



$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} \;-\; \mu \underline{\underline{\nabla_w}} Q(\boldsymbol{h}_n^{\mathsf{T}} \boldsymbol{w}_{n-1}, \boldsymbol{\gamma}(n))$$

$$\mu(n) = 1/n$$

gradient noise, $R$



# Performance Measures

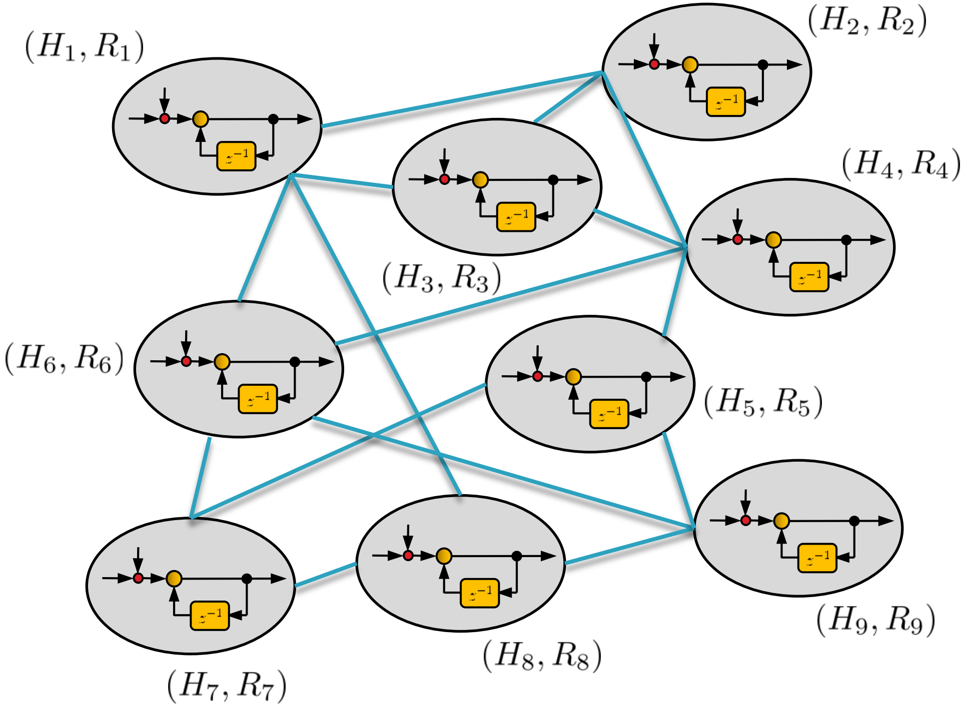


**Examples:**

$$\begin{cases} \limsup\limits_{n\to\infty} \mathbb{E}\,\|w^o - \boldsymbol{w}_n\|^2 & \text{(Mean-square-deviation)} \\ \limsup\limits_{n\to\infty} \mathbb{E}\,\left(J(\boldsymbol{w}_n) - J(w^o)\right) & \text{(Excess-risk))} \\ \text{Prob(error)} & \text{(Probability of error)} \end{cases}$$

# Networked Agents → Coupling






$$w^o \overset{\Delta}{=} \arg\min_{w} \sum_{k=1}^{N} J_k(w)$$

$$H_k \overset{\Delta}{=} \nabla_w^2 J_k(w^o)$$

$$R_k \overset{\Delta}{=} \text{gradient noise covariance}$$

# Influence of Topology?



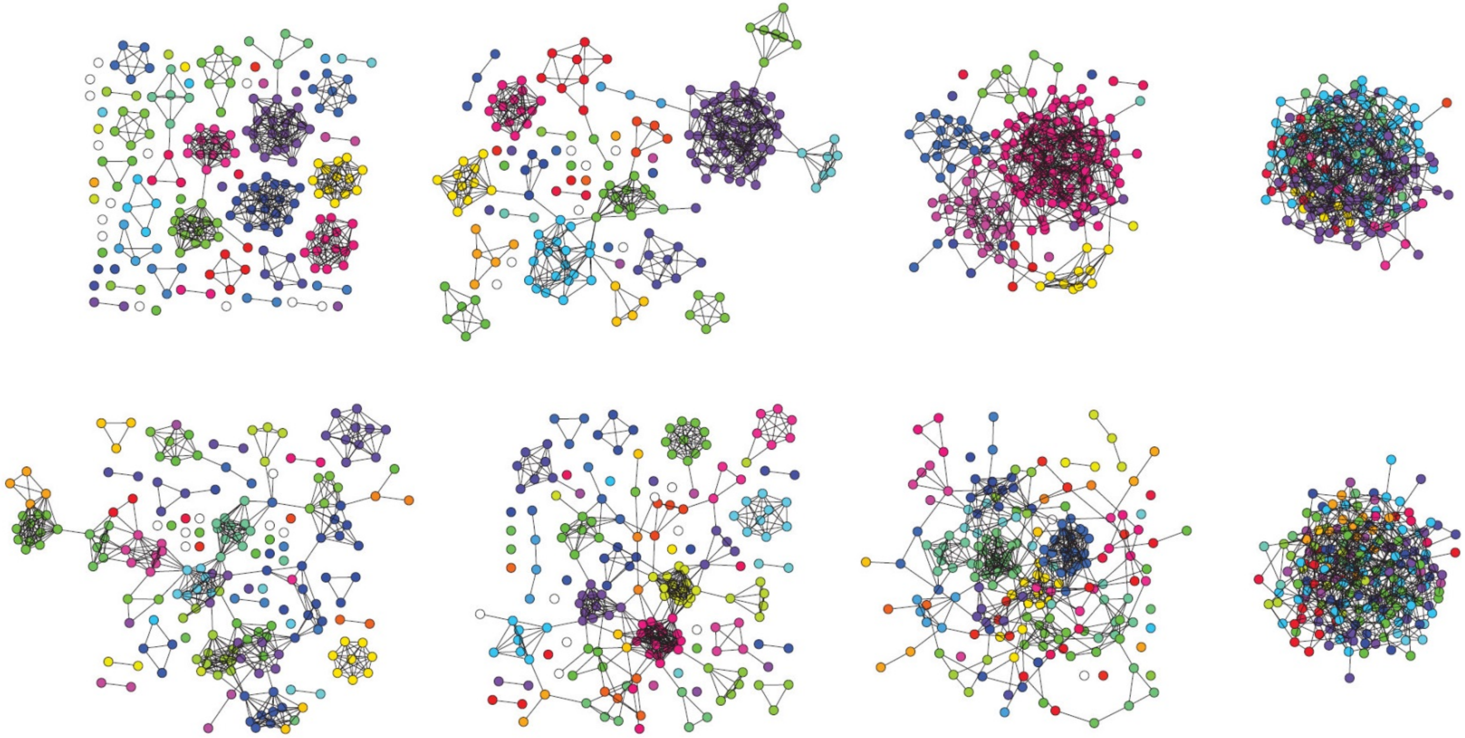


Image from *Agents, Interaction, and Complexity* Research Group website. University of Southampton.

# Some Relevant Questions



- Which topology has best performance?
- What aspects of the topology influence performance?
- Can different topologies deliver same performance?
- Is cooperation always beneficial?
- Can networks match performance of centralized solutions?

# Multi-Agent Network

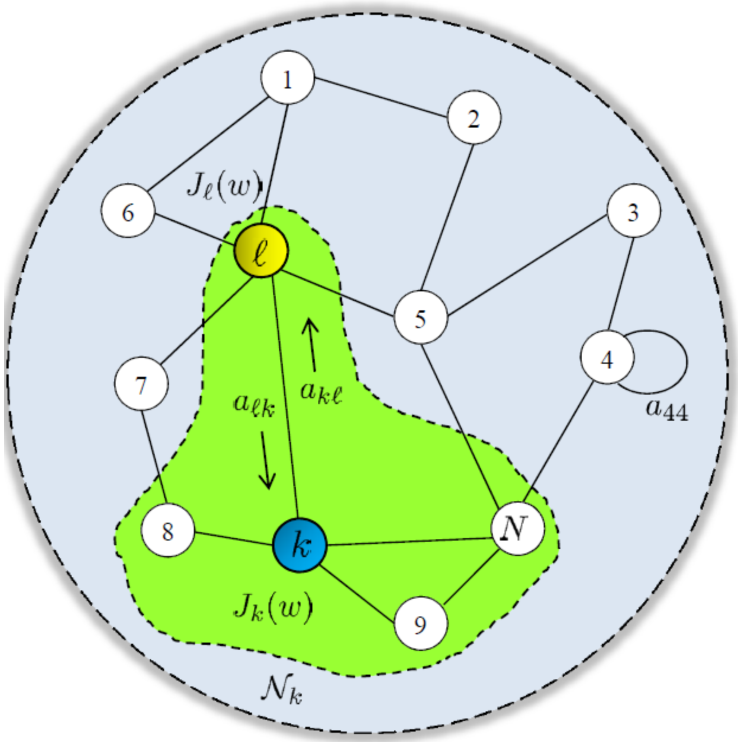


**Network with *N* vertices (agents):**

- Edges connecting agents.
- $a_{\ell k}$ : scales data from $\ell$ to $k$.
- $\mathcal{N}_k$ : neighbors of agent $k$.

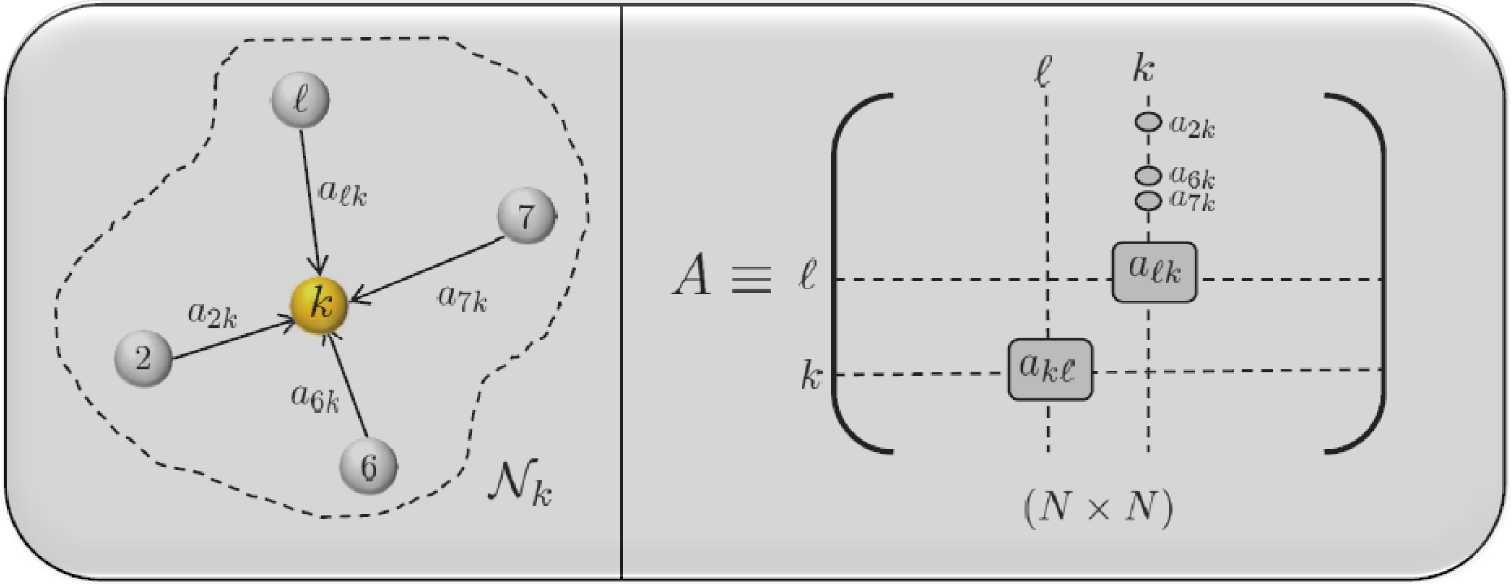
**Strongly-connected network:**

path with nonzero weights between any two agents plus at least one self-loop $(a_{kk} > 0)$.



# Combination Policy






$$a_{\ell k} \geq 0, \qquad \sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \qquad A^\mathsf{T}\mathbb{1} = \mathbb{1}$$ **(left-stochastic)**

# Perron-Frobenius Theorem



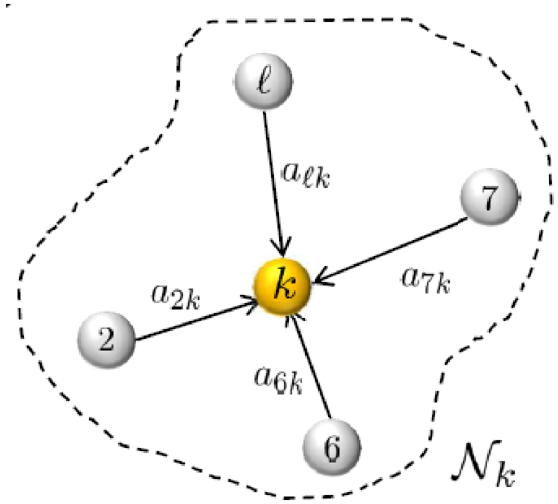
- $A$ has a single eigenvalue at one.
- All other eigenvalues are strictly inside the unit circle.
- **Perron vector:**

$$\boxed{Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0, \quad k = 1, 2, \ldots, N}$$

# Diffusion Learning



$$\begin{cases} \boldsymbol{\psi}_{k,n} &= \boldsymbol{w}_{k,n-1} - \mu \nabla_w Q_k(\boldsymbol{h}_{k,n}^{\mathsf{T}} \boldsymbol{w}_{k,n-1}, \boldsymbol{\gamma}_k(n)) \\ \boldsymbol{w}_{k,n} &= \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{\ell,n} \end{cases}$$




**Sayed et al.** (2006-2016): Lopes, Cattivelli, Tu, Zhao, Towfic, Chen, Yu, Vlaski, Ying.

# Diffusion Learning: *Variations*



$$\begin{cases} \boldsymbol{\psi}_{k,n} & = & \boldsymbol{w}_{k,n-1} - \mu \nabla_w Q_k(\boldsymbol{h}_{k,n}^{\mathsf{T}} \boldsymbol{w}_{k,n-1}, \boldsymbol{\gamma}_k(n)) \\ \boldsymbol{w}_{k,n} & = & \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{\ell,n} \end{cases}$$

Symmetry ensures stability regardless of topology!

**Variations exist to deal with non-smooth formulations:**

→ Sub-gradients

→ Proximal operators

→ Penalty-based

# Consensus Strategy



$$\boldsymbol{w}_{k,n} \;=\; \underline{\sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{w}_{\ell,n-1}} - \frac{\tau}{n} \nabla_w Q_k(\underline{\boldsymbol{h}_{k,n}^{\mathsf{T}} \boldsymbol{w}_{k,n-1}}, \gamma_k(n))$$

*Decaying step-size*

*asymmetry*

**Tsitsiklis and Athans** (1984)
**Boyd and Xiao** (2004)
**Moura et al** (2009)

# Consensus Strategy



$$\boldsymbol{w}_{k,n} \quad = \quad \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{w}_{\ell,n-1} - \frac{\tau}{n} \nabla_w Q_k(\boldsymbol{h}_{k,n}^{\mathsf{T}} \boldsymbol{w}_{k,n-1}, \gamma_k(n))$$

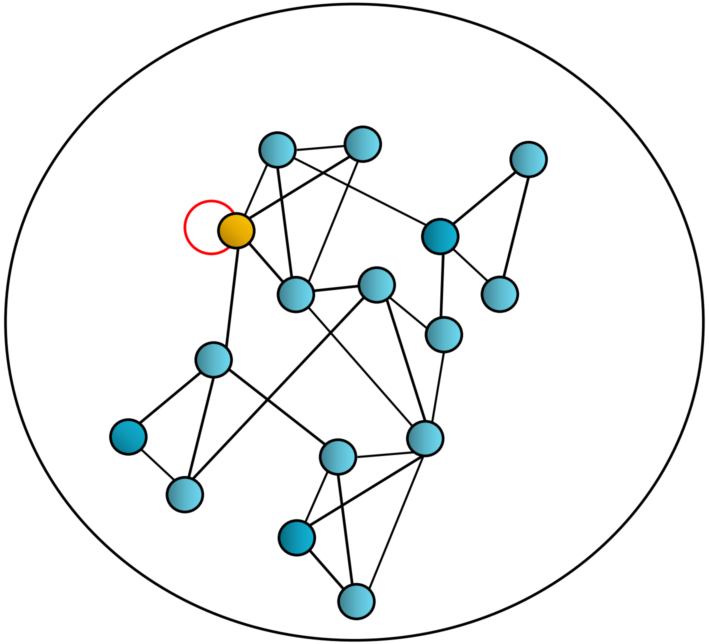
- **Solves well optimization problems.**
- **Problematic for adaptation and learning:**

  Asymmetry → **causes instability**

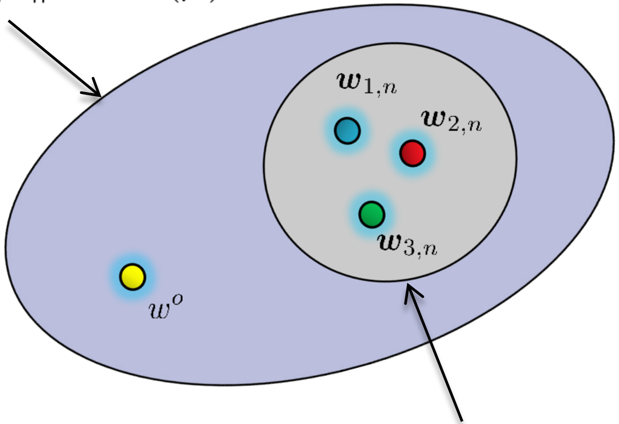
  Diminishing step-sizes → **limit adaptation & learning**

# THEOREM A: *Stability & Agreement*





$$\limsup_{n\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,n}\|^2 = O(\mu)$$

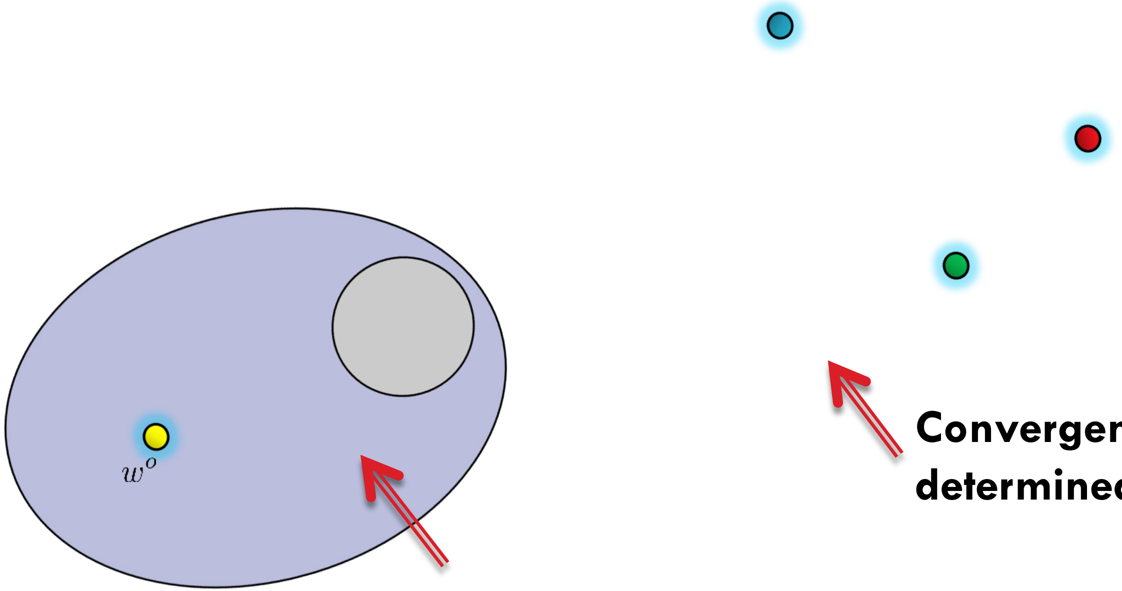


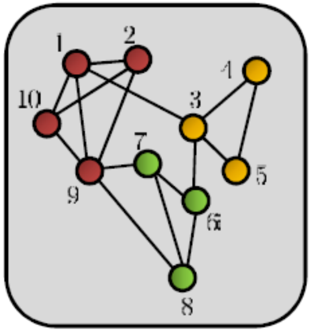

$$\limsup_{n\to\infty} \mathbb{E}\|\boldsymbol{w}_{k,n} - \boldsymbol{w}_{\ell,n}\|^2 = o(\mu)$$

**Zhao & Sayed** (*IEEE Trans. Signal Process.*, 2015)

# Two (not one) Rates of Convergence!





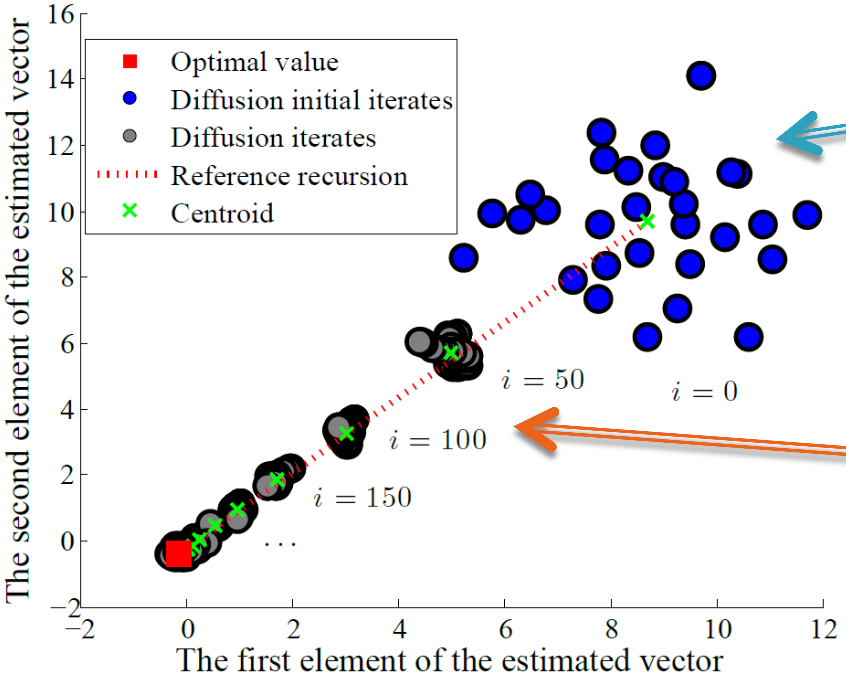

**Convergence rate determined by** $|\lambda_2(A)|$.

**A second convergence rate determined by** $\alpha \in (0, 1)$.

**Chen & Sayed** (*IEEE Trans. Infor. Thy.*, 2015)

# Two (not one) Rates of Convergence!






**Agents catch up with centralized solution at rate $|\lambda_2(A)|$.**

**Agents approach steady-state at a second rate $\alpha$.**

**Chen & Sayed** (*IEEE Trans. Infor. Thy.*, 2015)

# THEOREM B: Scaling Law #1



**For sufficiently small step-sizes:**

$$\limsup_{n\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,n}\|^2 \;=\; \frac{\mu}{2}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} p_k H_k\right)^{-1}\left(\sum_{k=1}^{N} p_k^2 R_k\right)\right] + o(\mu)$$

$$\alpha \;=\; 1 - 2\mu\lambda_{\min}\left(\sum_{k=1}^{N} p_k H_k\right) + o\left(\mu\right)$$

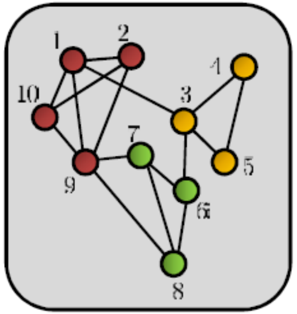
convergence rate

**Zhao & Sayed** (*IEEE Trans. Signal Process.*, 2012)
**Chen & Sayed** (*IEEE Trans. Infor. Thy.*, 2015)

# Network of Learners

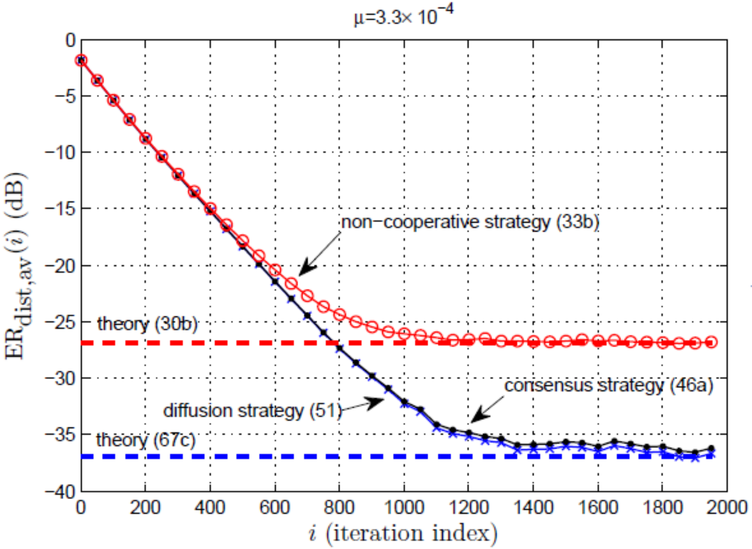


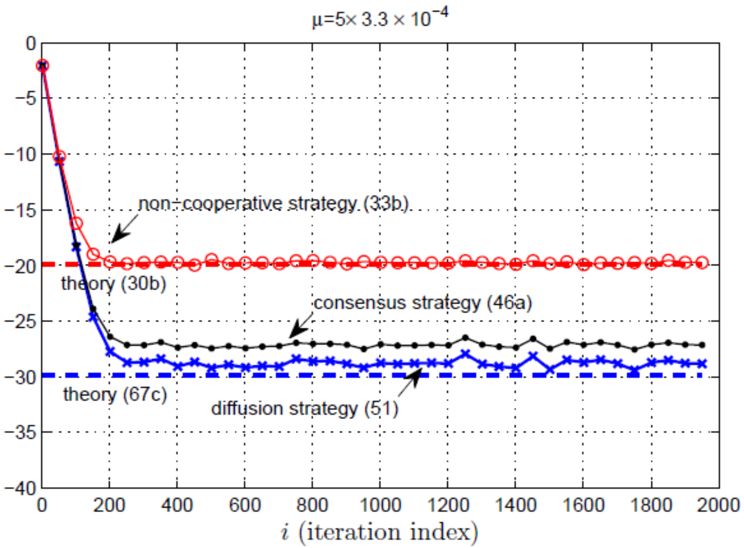
network of learners




(Metropolis rule)

$$\rho = 10$$



# Optimizing the Topology



$$A^o \;\stackrel{\Delta}{=}\; \arg\min_{A} \operatorname{Tr}\left(\sum_{k=1}^{N} p_k^2 H^{-1} R_k\right)$$
$$\text{subject to } \; Ap = p, \; \mathbb{1}^{\mathsf{T}} p = 1, \; p_k > 0$$

$$\theta_k^2 \;\stackrel{\Delta}{=}\; \operatorname{Tr}(H^{-1} R_k)$$

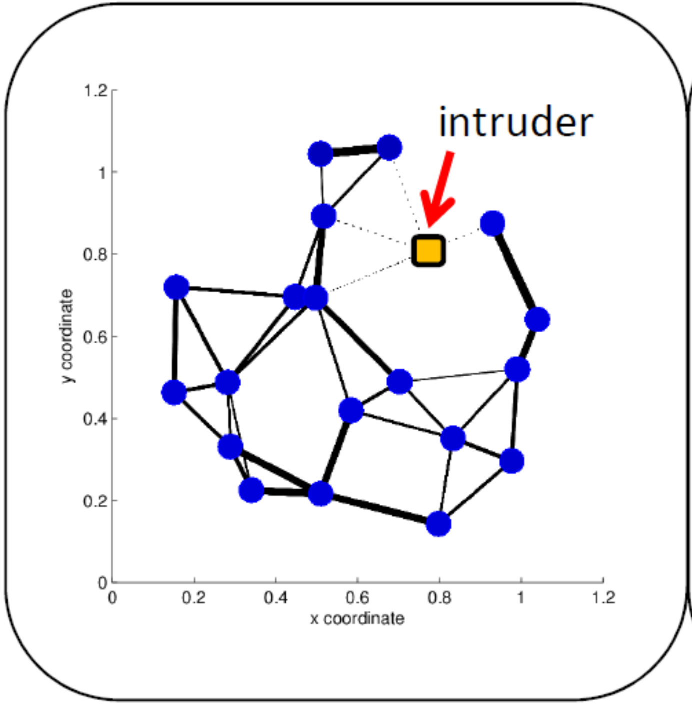
**(1/SNR measure)**

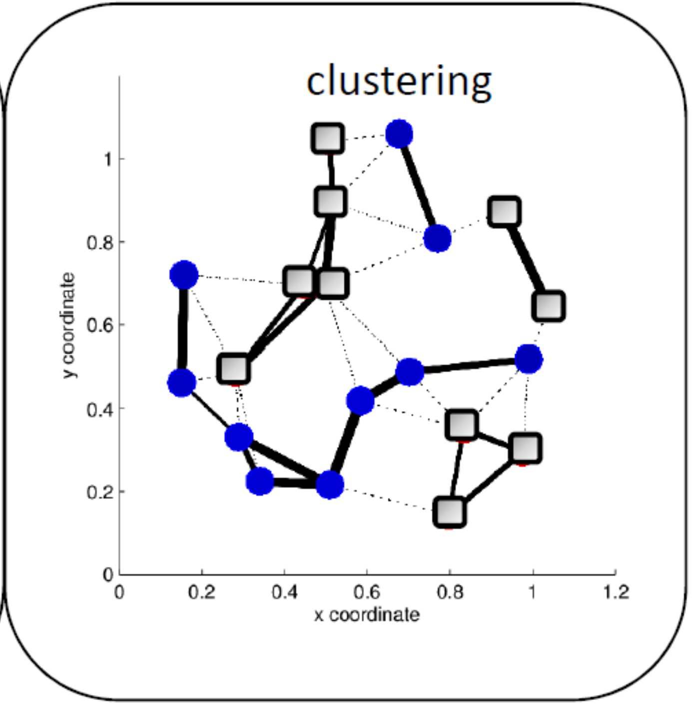
$$a_{\ell k}^o = \frac{\theta_k^2}{\max\left\{n_k \theta_k^2,\; n_\ell \theta_\ell^2\right\}}$$

**(Hastings rule)**

# Application: ***Detecting Intruders***



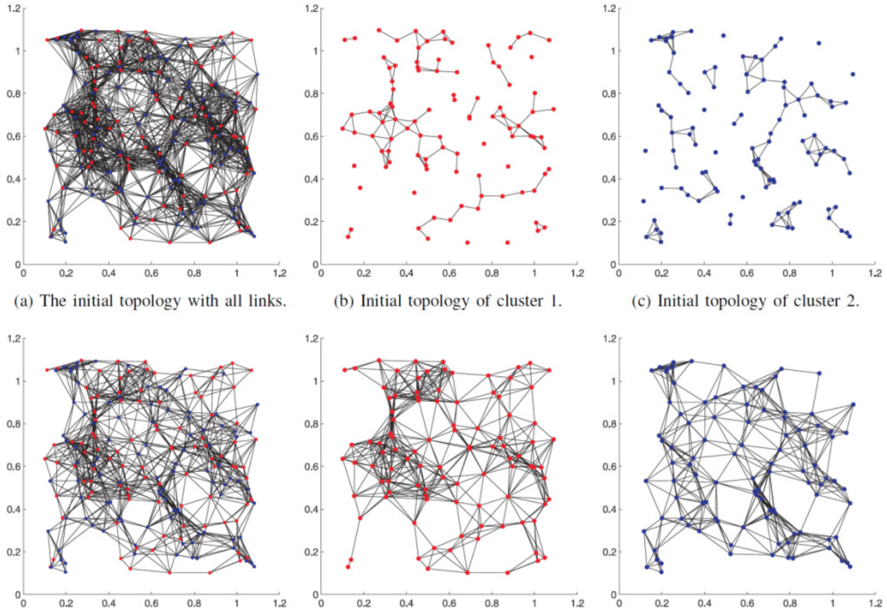



**Zhao & Sayed** (*IEEE Trans. Signal Process.*, 2012)

# Application: *Clustering*





(a) The initial topology with all links. (b) Initial topology of cluster 1. (c) Initial topology of cluster 2.

(d) The final topology at steady-state. (e) Resulting topology of cluster 1. (f) Resulting topology of cluster 2.

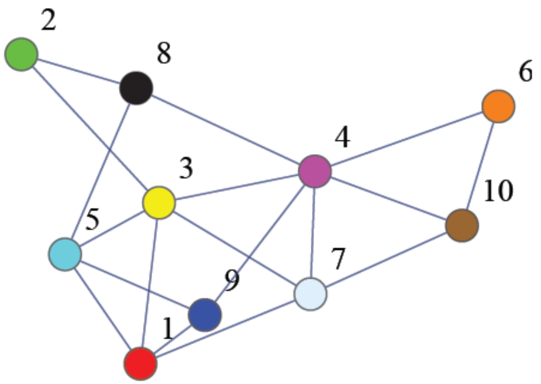
**Zhao & Sayed** (*IEEE Trans. Signal Process.*, 2015)

# THEOREM C: Scaling Law #2



Binary state of nature represented by hypotheses $\mathcal{H}_o$ and $\mathcal{H}_1$.

Using large deviation theory and exact asymptotic analysis, as $n \to \infty$.




(false alarm or mis-detection)

$$\text{Prob}_k[\text{error}] \;=\; e^{-(1/\mu)[a+o(1)]}$$

**Matta, Braca, Marano & Sayed** (*IEEE Trans. Infor. Thy*, 2016)

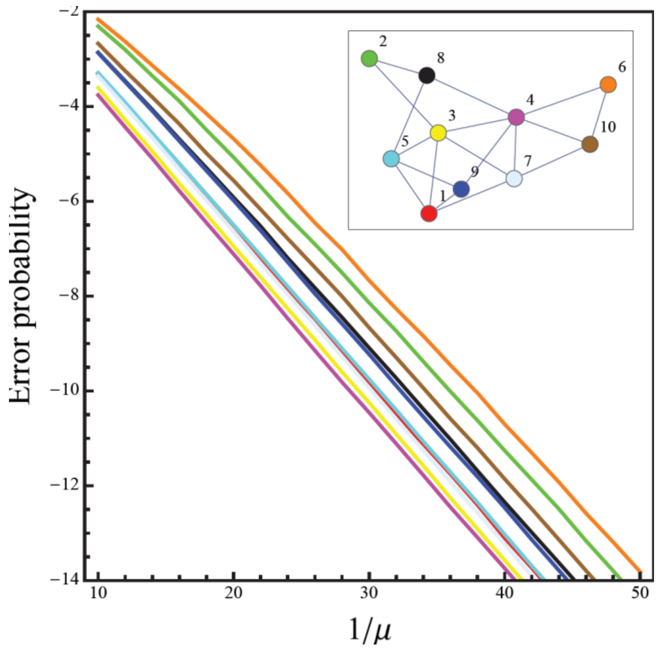
# Inverse Modeling



$$\mathrm{Prob}_k[\text{error}] \;=\; e^{-(1/\mu)[a+o(1)]}$$

**IMPLICATIONS:**

- Error probabilities vanish exponentially with $1/\mu$.
- Same exponent $a$: parallel curves
- Connectivity matters → o(*1*)




*More peripheral agents perform worse.*

# Fundamental Tradeoff



| | Networked agents | Centralized inference with *N* iid obs. | Rate-constrained multi-terminal inference (CEO) |
|---|---|---|---|
| **MSE (distortion)** | $\propto \mu$ | $\propto 1/N$ | $\propto 1/R$ |
| **Error probability** | $\propto e^{-1/\mu}$ | $\propto e^{-N}$ | $\propto e^{-R}$ |

Reducing $\mu$ plays a role similar to increasing $(N, R)$.

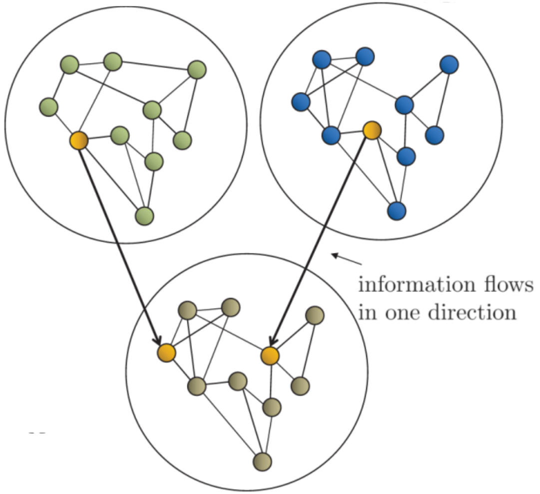
→ Interpretation: $\mu$ quantifies the cost of information.

# What about Weak Graphs?



**Weak connectedness can arise as a result of:**

- ✓ Intruder attacks by malicious agents;
- ✓ Failure by some critical links;
- ✓ Presence of stubborn agents;
- ✓ Information control;
- ✓ Asymmetric information dissemination over social platforms.

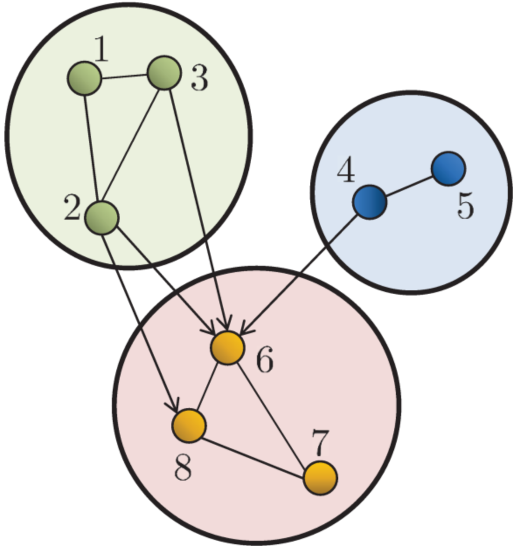


# Example



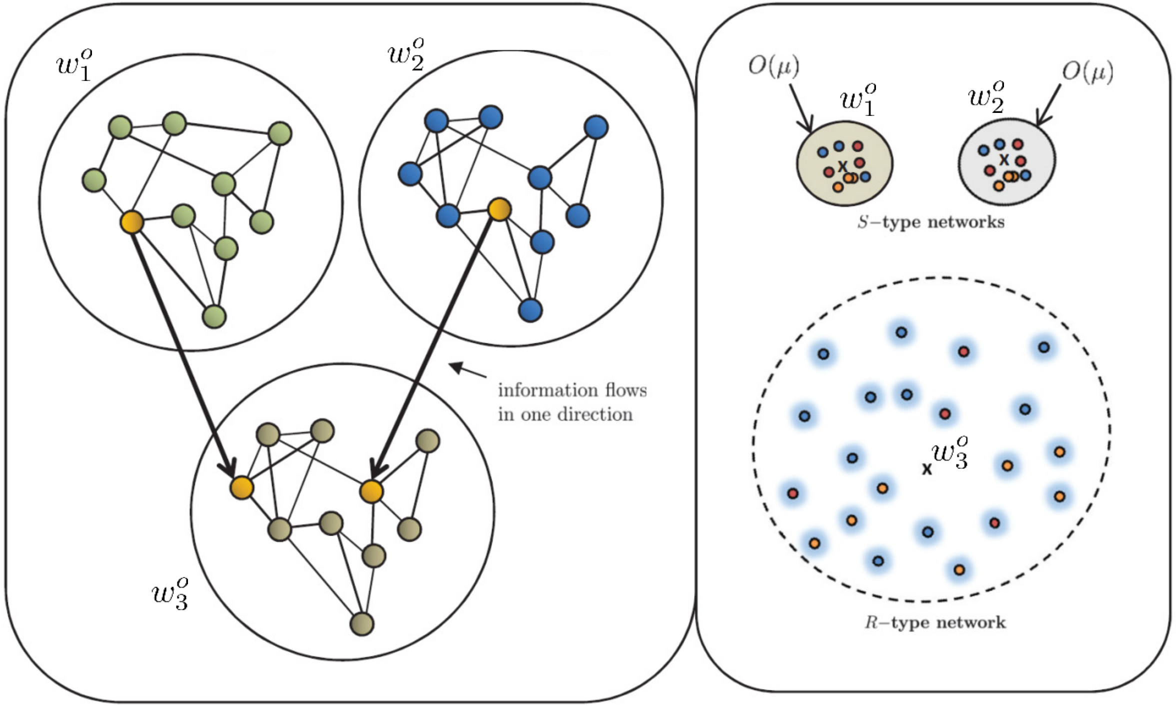
8 agents, S=2 sending sub-networks, and R=1 receiving sub-networks.

$$A = \begin{array}{c} \\ 1\\2\\3\\4\\5\\6\\7\\8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left[\begin{array}{ccccc|ccc} 0.2 & 0.2 & 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.4 & 0.1 & 0 & 0 & 0.2 & 0 & 0.4 \\ 0.3 & 0.4 & 0.1 & 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.3 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0.7 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0.2 & 0.3 & 0.2 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0.5 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.1 \end{array}\right] \end{array}$$



# Mind-Control, Defiance, Reconciliation






$$w^\bullet = \mathcal{W} w^o$$

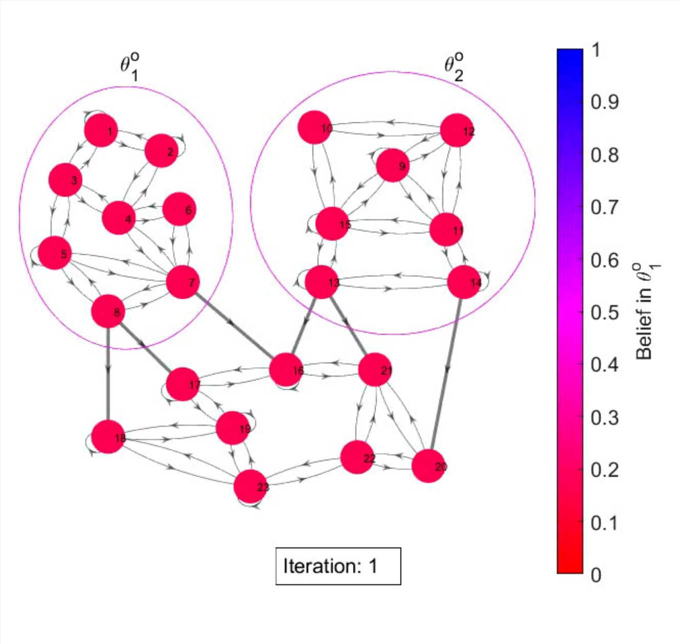
$$\mathrm{MSD}_{r,k} = \sum_{s=1}^{S} \alpha_k^2 \cdot \mathrm{MSD}_s$$

**Ying & Sayed** (*IEEE Trans. Infor. Theory*, 2015)
**Salami, Ying, & Sayed** (*Proc. ICASSP*, 2016)

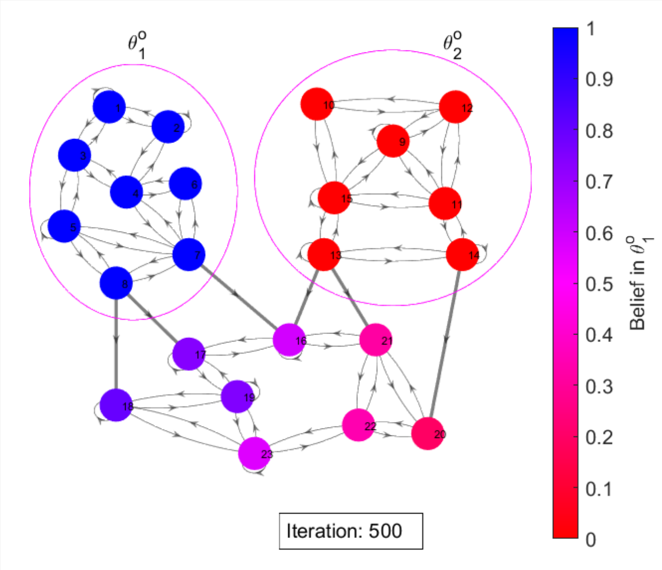
# Control Mechanism



**video animation**




**steady-state**



# Concluding Remarks



**Interesting phenomena arise when information is processed in a distributed manner over networks:**

Is more better?
Why is one topology better than the other?
How to ensure network stability?
How to adapt the topology and weights?
How to handle selfish agents? Intruders? Outliers?
Bio-inspired cognition. Swarming. Evasion procedures.

→ Many challenging open issues ….