Distributed Collaborative Non-Convex Optimization

A Case Study on ℓ-0 Constrained Graph Estimation

Mengdi Wang

ORFE, Princeton University

SIAM CSE, March 2015

Collaborators

Ethan X. Fang

Han Liu

Distributed Collaborative Optimization

Consider the social welfare maximization problem

$$\max_{x_i \in \mathcal{X}_i, \forall i} u_0\left(\frac{1}{N}\sum_{i=1}^{N} c_i(x_i)\right) + \frac{1}{N}\sum_{i=1}^{N} u_i(x_i)$$

- $u_i(x_i)$, $i = 1, \ldots, N$, is the private utility of the ith user
- u_0 is the social utility associated with some common goods (e.g., total energy consumption, total air pollution)
- $u_i, c_i, \mathcal{X}_i$ are nonconvex

Many decentralized problems have a natural near-separable structure.

Question:

Is there any intrinsic simplicity that we can leverage to design efficient distributed algorithms?

Application: Pricing of Public Resources

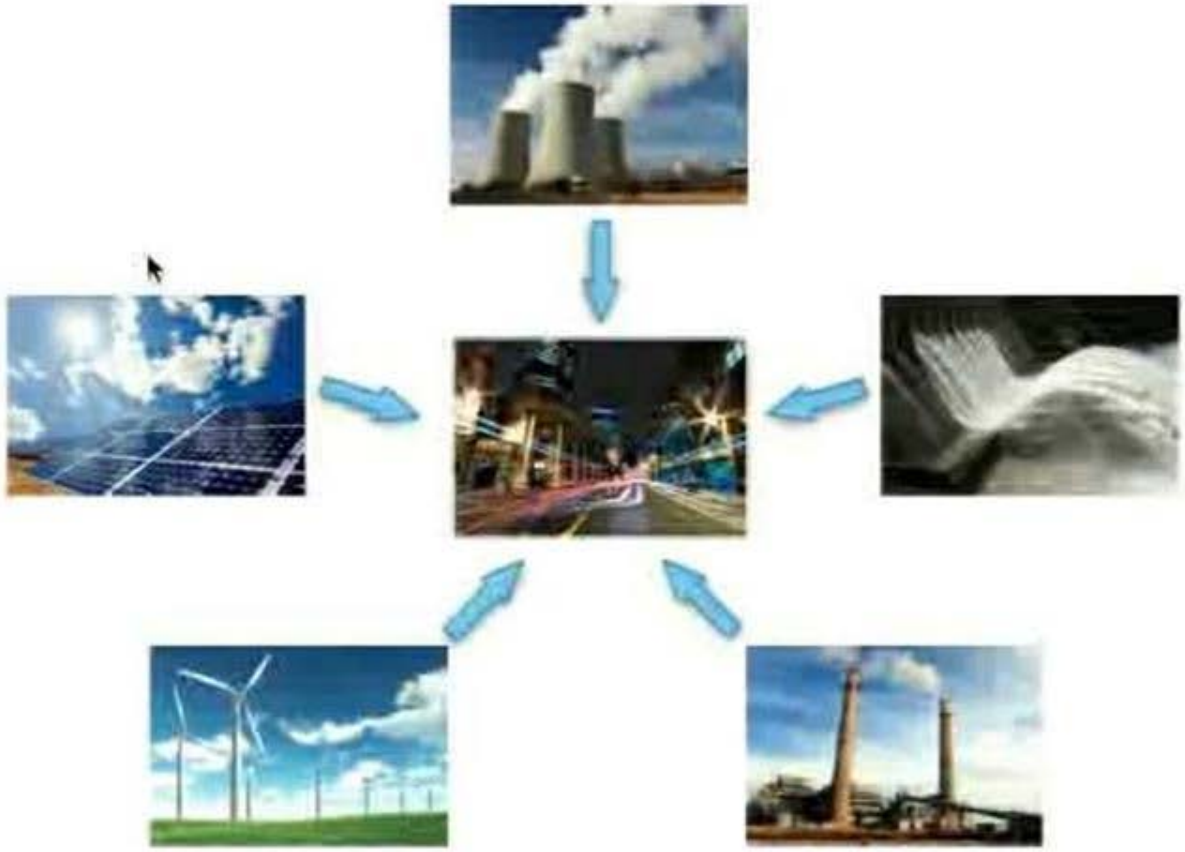


Figure : Individual users may have flexibility in selecting different energy usages, resulting in non-concave preference function.

Application: Graphical Models

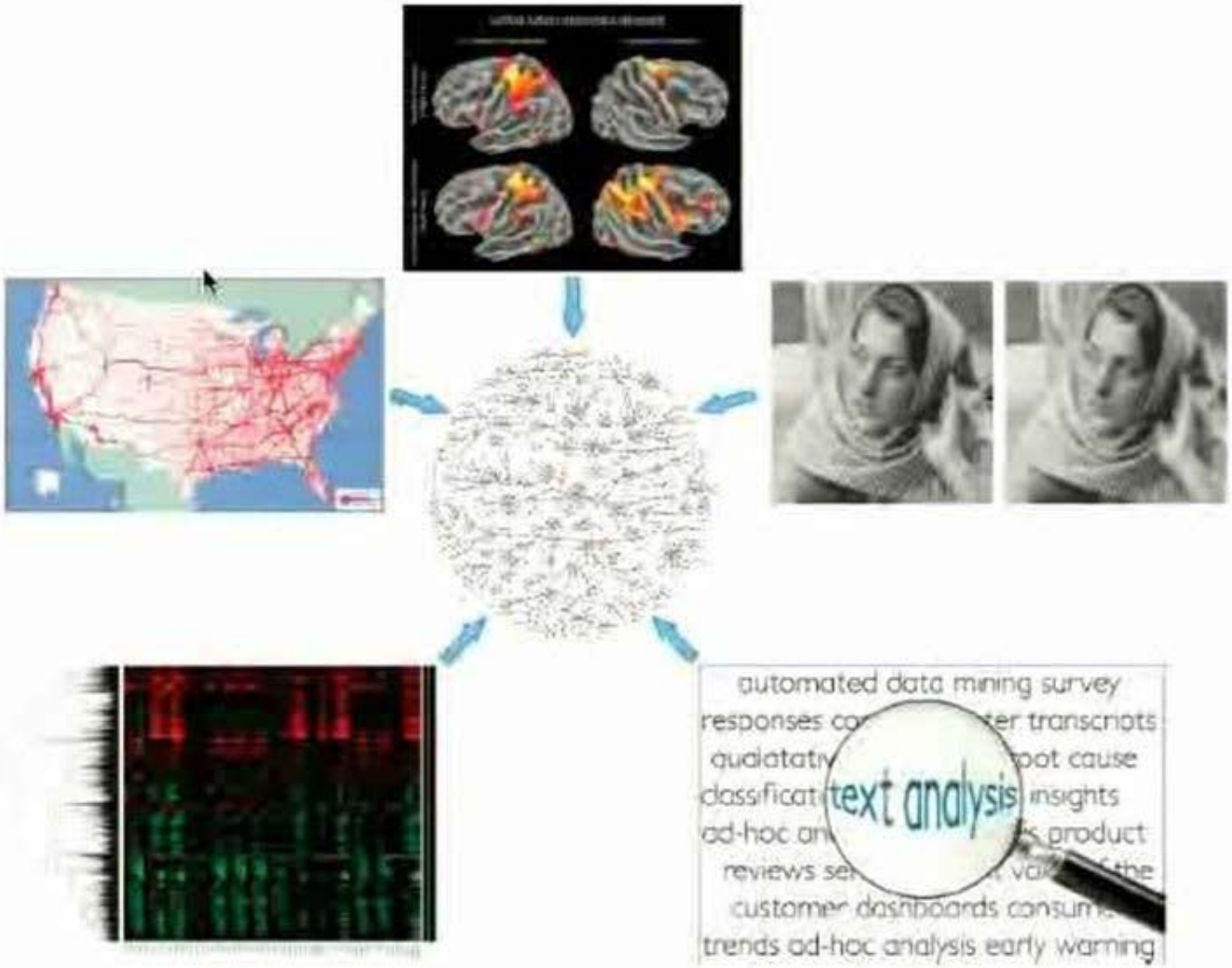
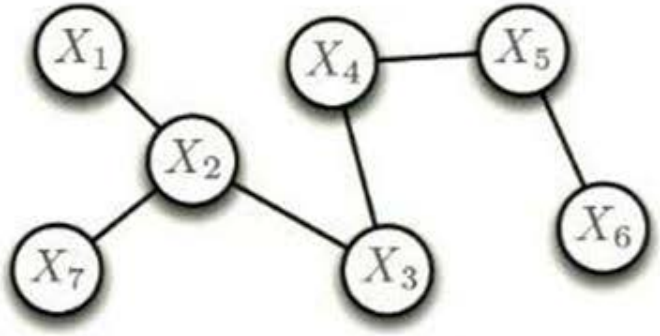


Figure: Graphical model is used in a wide range of applications.

Graphical Model

In the graph, each node represents a random variable. An edge is presented between X_j and X_k if and only if they are conditionally dependent given other nodes, i.e.,

$$\mathbb{P}(X_j = x_j, X_k = x_k | \boldsymbol{X}_{\setminus\{j,k\}}) = \mathbb{P}(X_j = x_j | \boldsymbol{X}_{\setminus\{j,k\}}) \mathbb{P}(X_k = x_k | \boldsymbol{X}_{\setminus\{j,k\}}).$$



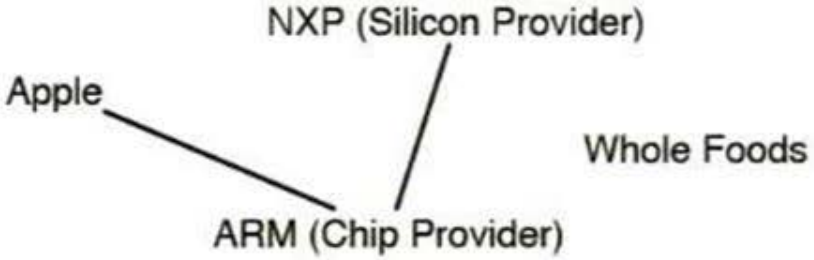
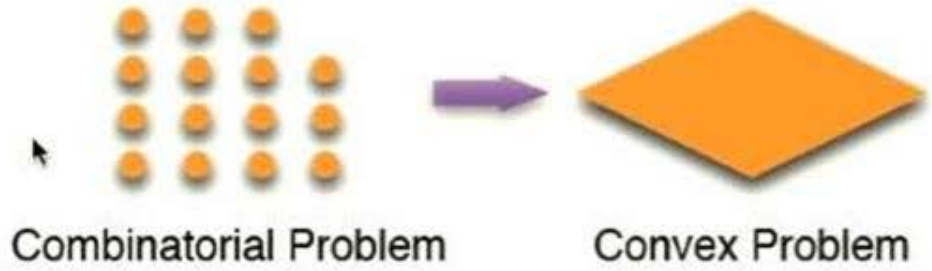


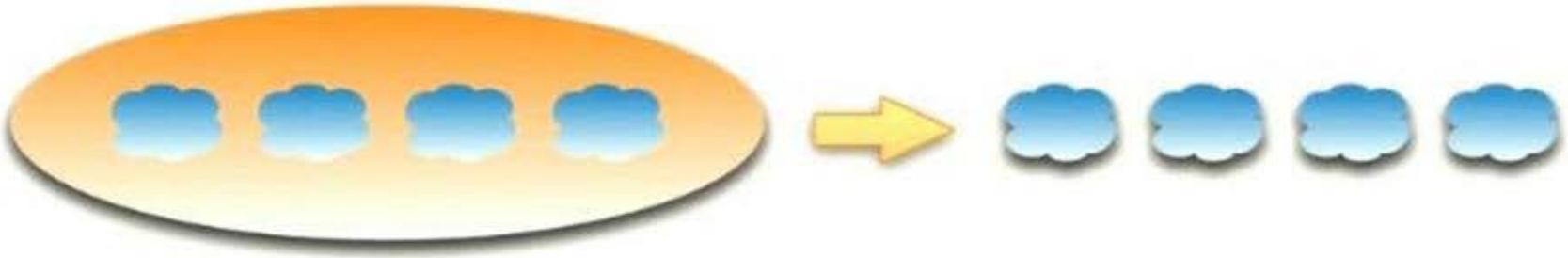
Figure : A class of graphical models explains the conditional independences.

Computational Challenge

This problem is combinatorial and NP-hard.



Convex relaxation would incur some unavoidable statistical errors.
In practice, each subproblem can be easy to solve. However, the whole problem is difficult due to the global sparsity constraint.



Sparse Graph Estimation

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d} \frac{1}{d}\sum_{j=1}^{d} \ell_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K.$$

Why ℓ-0?

We care about estimation error as well as computation efficiency.

- The ℓ-0 constraint exactly describes the statistic modeling assumptions:

  "The total number of arcs is bounded by $K \ll d^2$."

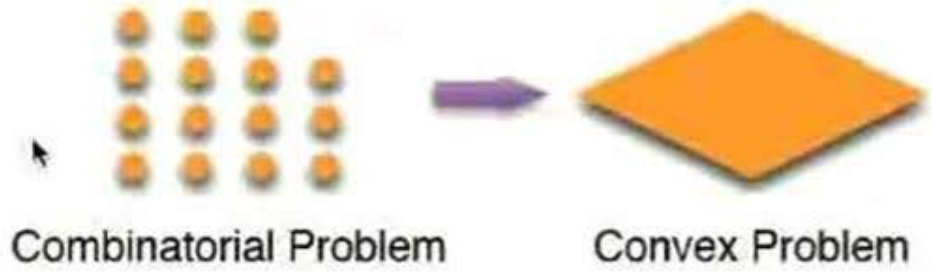
- Avoids the statistical estimation gap between ℓ_0 and ℓ_1 norm

Objective: Estimate the True Graph

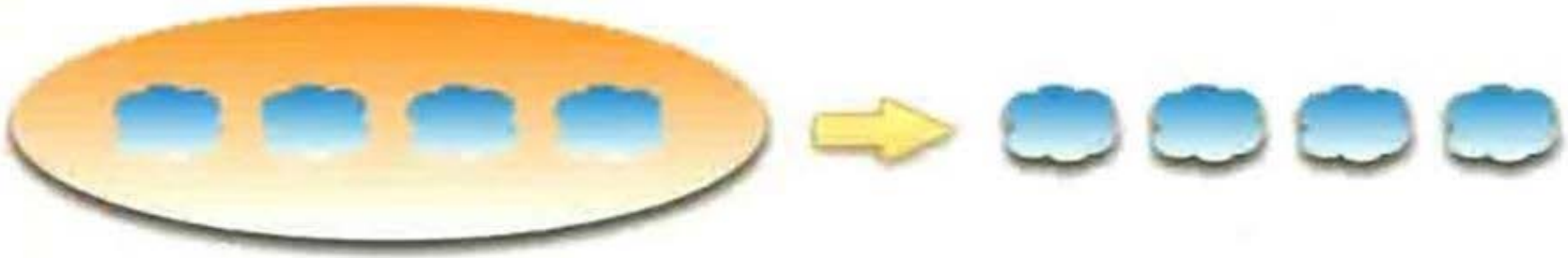
- Let β^* be the *true* coefficients (not the optimal solution to the optimization problem)
- We want to find an (approximate) solution $\widehat{\beta}$ to the optimization problem and use it as an estimator of β^*

Computational Challenge

This problem is combinatorial and NP-hard.



Convex relaxation would incur some unavoidable statistical errors.
In practice, each subproblem can be easy to solve. However, the whole problem is difficult due to the global sparsity constraint.



Sparse Graph Estimation

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d} \frac{1}{d}\sum_{j=1}^{d} \ell_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K.$$

Why ℓ-0?

We care about estimation error as well as computation efficiency.

- The ℓ-0 constraint exactly describes the statistic modeling assumptions:

  "The total number of arcs is bounded by $K \ll d^2$."

- Avoids the statistical estimation gap between ℓ_0 and ℓ_1 norm

Objective: Estimate the True Graph

- Let β^* be the *true* coefficients (not the optimal solution to the optimization problem)
- We want to find an (approximate) solution $\widehat{\beta}$ to the optimization problem and use it as an estimator of β^*

Intuition from (non-)Convex Geometry

The intuition traces way back to Shapley-Folkman Theorem that

"The sum of many nonconvex sets is close to being convex."

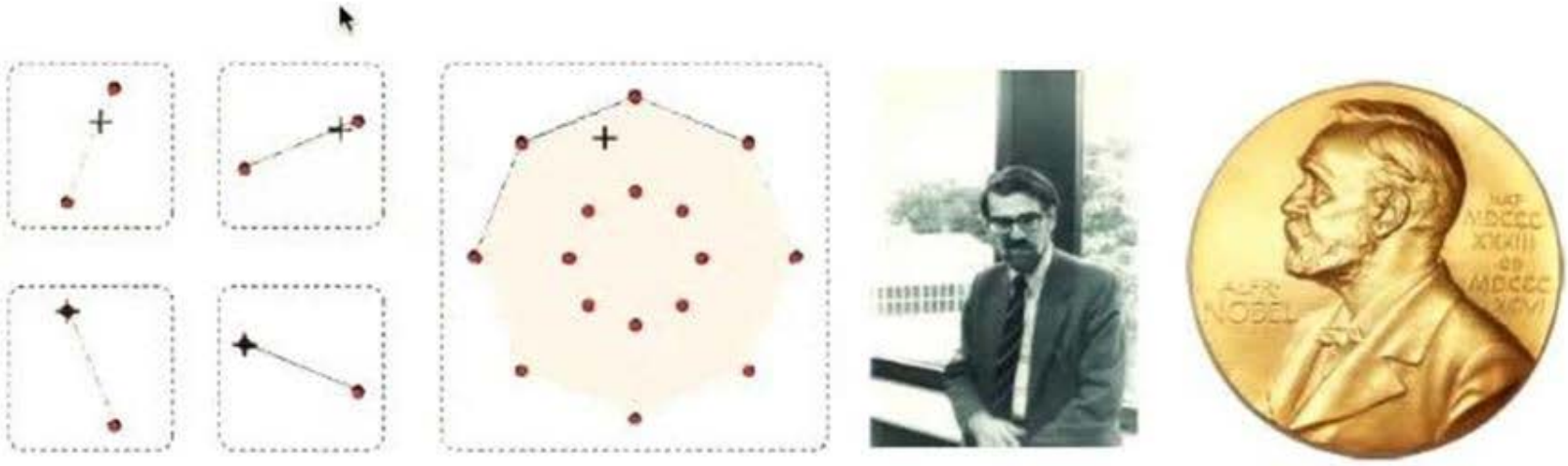


Figure : Shapley-Folkman Theorem reveals the insight that the sum of many nonconvex sets is close to being convex. Lloyd Shapley, 2012 Nobel Economics Laureate.

Blessing of Large Scales

We use the dual solution $\widehat{\beta}_j$'s as the estimators. Let β_j^*'s denote the truces and n denote sample size. We get

$$\frac{1}{d}\sum_{j=1}^{d}\|\widehat{\beta}_j - \beta_j^*\|_2^2 = \underbrace{\mathcal{O}\Big(\frac{(K/d)\log d}{n}\Big)}_{\text{unavoidable}} + \underbrace{\mathcal{O}\Big(\frac{\delta_n}{d}\Big)}_{\text{duality gap}}.$$

- $\mathcal{O}\Big(\frac{(K/d)\log d}{n}\Big)$ matches the information-theoretic lower bound.
- $\mathcal{O}(\frac{\delta_n}{d})$ is the duality gap (difference between ℓ_0 problem and its dual) vanishes as graph size increases.
- As $n \to \infty$, $\delta_n \to \max_{j=1,\ldots,d} D_{KL}\Big(\mathbb{P}(\boldsymbol{X}_j)\Big\|\mathbb{P}\mathbb{P}(\boldsymbol{X}_j \mid \boldsymbol{X}_{\setminus j})\Big)$ The duality gap constant converges to the maximum Kullback-Leibler divergence between the marginal distribution of a node j and its conditional distribution.
- For large graphs:

duality gap $\ll$ unavoidable statistical uncertainty

Brain Voxel Network Estimation

Figure: Understanding brain network is an initial step to understand our brain.

fMRI Data

The fMRI (Functional Magnetic Resonance Imaging) technology measures brain activity by detecting associated changes in blood flow. Number of voxels is $d \geq 30,000$.

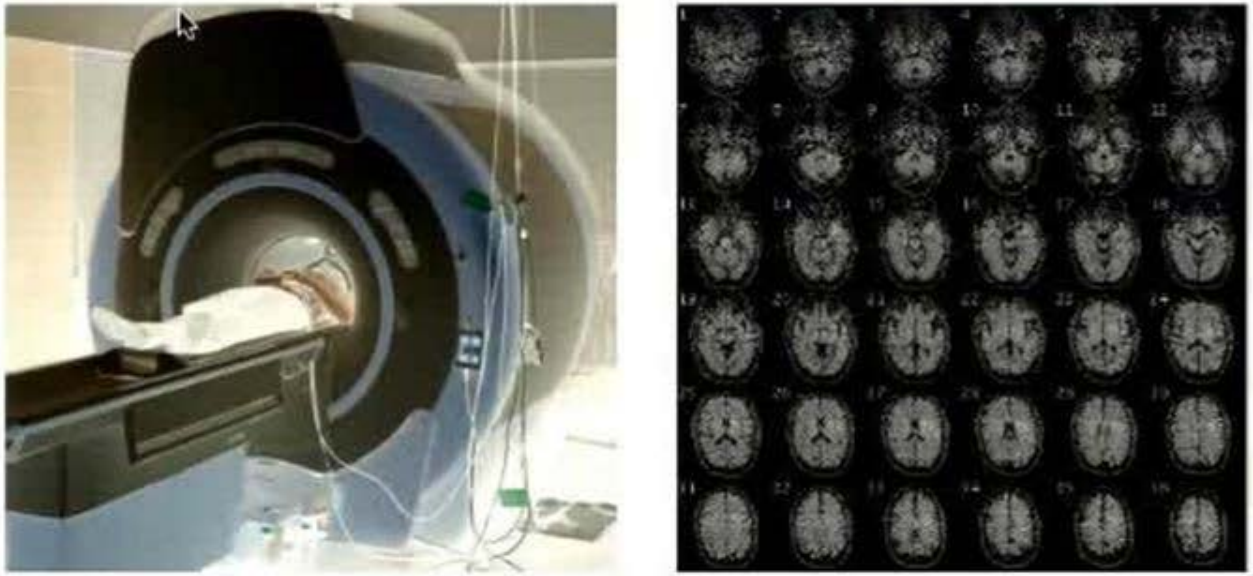


Figure : fMRI data.

Real Data Result

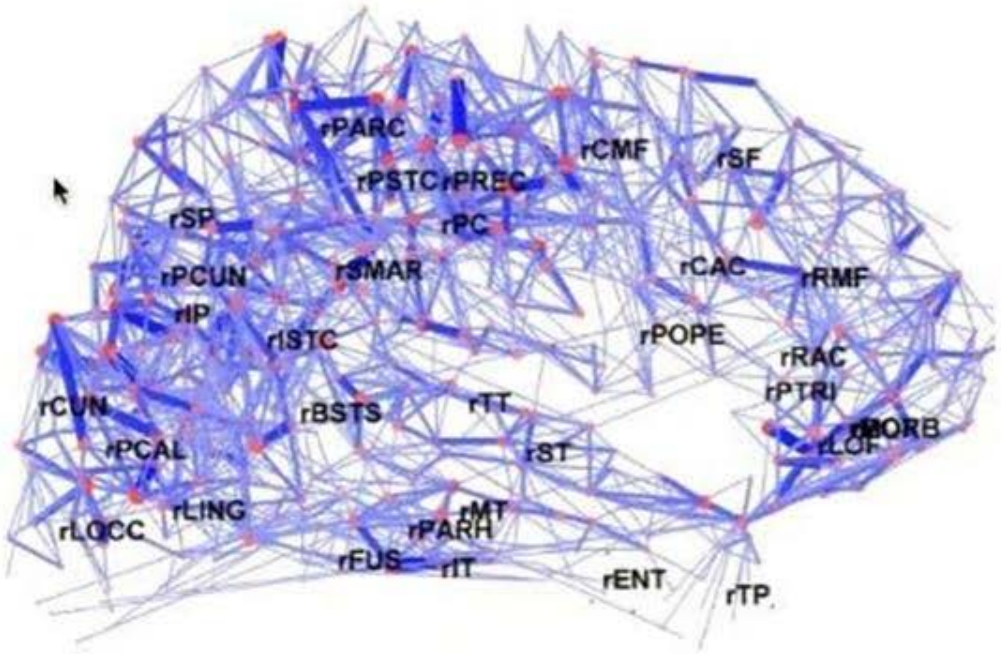


Figure : The estimated graph reveals that voxels related to a same function are highly connected.

More Generally: Distributed Collaborative Optimization

$$\max_{x_i \in \mathcal{X}_i, \forall i} u_0\left(\frac{1}{N}\sum_{i=1}^{N} c_i(x_i)\right) + \frac{1}{N}\sum_{i=1}^{N} u_i(x_i)$$

- u_0 is the convex social utility function
- u_i is the individual utility of the ith player; $u_0, u_i, c_i, \mathcal{X}_i$ are nonconvex

Fenchel Dual Problem

$$\inf_{\lambda \in \Re^m} \left\{ -u_0^\star(\mu) + \frac{1}{N}\sum_{i=1}^{N} \sup_{x_i \in \mathcal{X}_i} \{\lambda' c_i(x_i) + u_i(x_i)\} \right\}.$$

Duality Gap Results

- The duality gap vanishes to 0 as $N \to \infty$
- There exists a dual solution (out of many bad dual solutions) that achieves the duality gap.

Decentralized Coordination Algorithm: A Sketch

- At time t, the current price vector is λ_t.
- Under an incentive mechanism, each user reports multiple solutions to the penalized problem

$$\max_{x_i \in \mathcal{X}_i} u_i(x_i) - \lambda_t' c_i(u_i)$$

  and receives a reward for providing flexibility.
- The central decision maker chooses an admissible solution for each user, and updates the price vector from to λ_{t+1}.

Dynamic Convergence to Social Optimum

- The iteration converges to an ϵ-social optimum with $\epsilon =$ duality gap.
- If users are willing to report their utility values, there exists an algorithm achieving regret

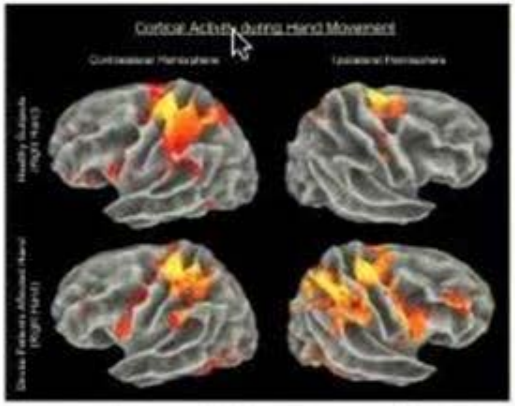
$$\mathcal{O}(\log T + \epsilon T)$$

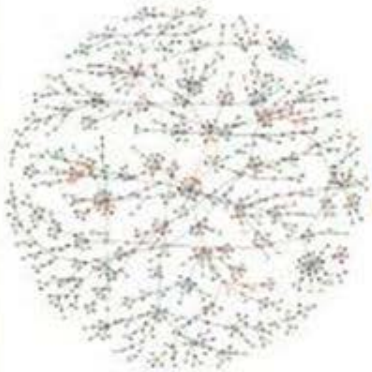
- If users don't report their utility values, there is an algorithm achieving regret

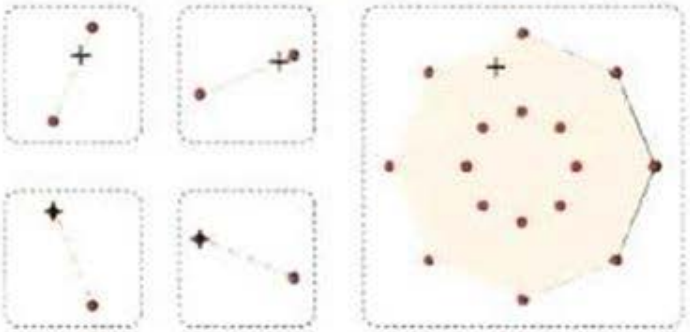
$$\mathcal{O}(\sqrt{T} + \epsilon T)$$

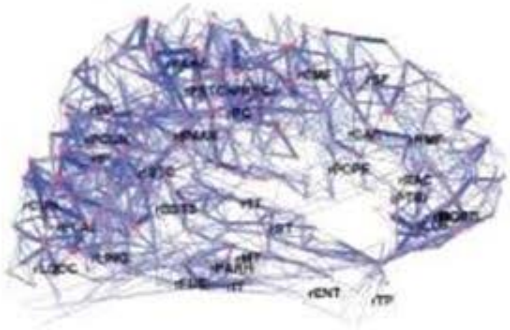
Summary

$$\min_{\beta_1,\ldots,\beta_d} \frac{1}{d}\sum_{j=1}^{d} \ell_j(\beta_j), \text{ subject to } \sum_{j=1}^{d} g_j(\beta_j) \leq \mathrm{b}.$$









Vanishing duality gap in large distributed systems.