Information Entropy-Based Node Sampling for Communication-Efficient Decentralized Learning

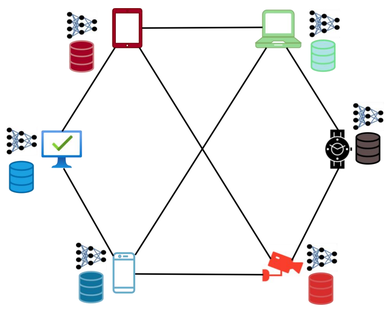
Jaiprakash Nagar[a], Zheng Chen[b], Photios A. Stavrou[a]

[a]The Foundation & Algorithm Group, Communication Systems Department, EURECOM, France

[b]Department of Electrical Engineering, Linkoping University

Motivations

Training a global machine learning model in decentralized settings requires networked nodes to exchange model parameters regularly with their neighbors to achieve a learning algorithm's convergence. This frequent information exchange among networked nodes results in huge communication overheads. Thus, communication cost becomes a bottleneck for wireless networks facing resource scarcity in terms of limited bandwidth and energy. The communication cost in terms of the number of transmission slots per communication round can be reduced by exploiting graph sparsification for communication-efficient decentralized learning, as ”not all links are equally important in a graph”. This translates to having the more important links activated more often than the less important links to achieve convergence. Further, it is also the fact that ” not all nodes are equally important in a graph”; therefore, more critical nodes should communicate more frequently than the trivial nodes.



Decentralized machine learning scenario.

- Different methods exist to compute node importance in a connected graph, where betweenness centrality methods provide more accurate node ranking in terms of their importance than any other centrality method.
- The betweenness centrality method does not capture crucial nodes in densely connected topologies.
- Also, betweenness centrality is unsuitable for computing node importance for irregular topologies as the BC value of a degree one node is zero.

Problem Formulation

- The optimization objective is to find parameter vector $x \in \mathbb{R}^d$ that minimizes the global loss function $F(x)$

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} F_i(x)$$

where $F_i(x) = \frac{1}{\mathcal{S}_i}\sum_{s\in\mathcal{S}_i} \ell(x;s)$

Decentralized Stochastic Gradient Descent (D-SGD)

- Every iteration of the D-SGD algorithm consists of two steps:
  1. Stochastic gradient update: $x_i^{\left(k+\frac{1}{2}\right)} = x_i^{(k)} - \gamma g_i^{(k)}$

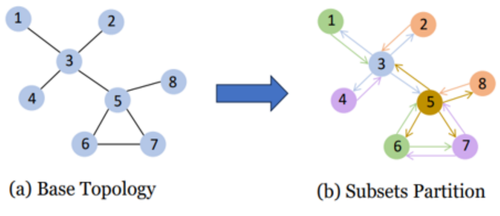
     e.g., $g_i^{(k)} = \nabla F_i\left(x_i^{(k)}; \xi_i^{(k)}\right)$, with $\xi_i^{(k)} \subseteq \mathcal{S}_i$ being a randomly sampled mini-batch of data
  2. Consensus averaging: $x_i^{(k+1)} = \sum_{j=1}^{N} W_{ij}\, x_j^{\left(k+\frac{1}{2}\right)}$ Mixing Matrix: $W(k) = \left[W_{ij}(k)\right]_{i,j\in[N]}$

Mixing Matrix Properties and Design

- Distributed averaging at iteration step t: $x(k+1) = \mathrm{W}x(k)$
- Necessary and sufficient conditions for the convergence to $\frac{1}{N}11^T x(0)$
  - Each row/column sum up to one: $W1 = 1, 1^T W = 1^T$
  - Spectral radius: $\rho(W - J) < 1$
- Common choice of mixing matrix: $\mathrm{W}(k) = I - \alpha L(k)$

  where $L(k) = D(k) - A(k)$ symmetric Laplacian Matrix of a graph

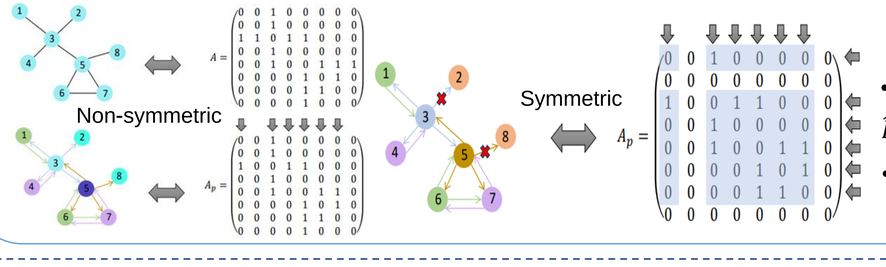
Base Topology Partition into Collision-free Sets



Base topology divided into q subsets:

- $S_i = (V_i, E_i)$
- $\cup_{i=1}^{q} V_i = V$
- $V_i \cap V_j = \emptyset, \forall i \neq j$
- $\cup_{i=1}^{q} E_i = E$
- $E_i \cap E_j = \emptyset, \forall i \neq j$

Node Scheduling: Sampling Columns of Adjacency Matrix



$A_p = \hat{A}(k) = Q(k)AQ(k)$

- Laplacian in k-th iteration $\hat{L}(k) = diag\left(\hat{A}(k)u\right) - \hat{A}(k)$
- Related mixing matrix is $\mathrm{W}(k) = I - \alpha\hat{L}(k)$

Information Entropy-Based Node Importance

- Probability of an edge connecting node v_i and v_j with degree d_i and d_j.

$$P(v_i, v_j) = \frac{1}{d_i d_j}$$

- Compute self-information of the edge $\mathcal{W}(v_i, v_j)$ connecting node v_i and v_j.

$$\mathcal{W}(v_i, v_j) = -\log_2 P(v_i, v_j) = \log_2(d_i d_j)$$

- Compute $\mathcal{W}(v_i)$ representing the sum of self-information of edges with v_i as one of its endpoints.

$$\mathcal{W}(v_i) = \sum_{v_j \epsilon \psi(v_i)} \mathcal{W}(v_i, v_j) \qquad N(v_i)\text{: Set of neighbors of node } v_i$$

- Compute $\mathcal{W}^+(v_i)$ representing the sum of self-information of edges that v_i and its neighbors are one endpoint of these edges.

$$\mathcal{W}^+(v_i) = \sum_{v_j \epsilon \psi(v_i)} \mathcal{W}(v_j) \quad \text{where, } \psi(v_i) = N(v_i) \cup \{v_i\}$$

- Probability corresponding to node v_j denoted by $P(v_j)$ is computed as

$$P(v_j) = \frac{\mathcal{W}(v_j)}{\mathcal{W}^+(v_i)} \qquad \text{Such that } \sum_{v_j \epsilon \psi(v_i)} \frac{\mathcal{W}(v_j)}{\mathcal{W}^+(v_i)} = 1$$

- Information entropy of node v_i

$$E(v_i) = -\sum_{v_j \epsilon \psi(v_i)} P(v_j)\log_2 P(v_j)$$

Probabilistic Sampling of Collision-free Subsets

- Let b_i denotes the i^{th} node's importance such that $\sum_{i=1}^{N} b_i = 1$

$$\text{Subset importance } S_j\text{: } b_{S_j} = \sum_{i=1}^{N} b_i 1_{\{i\epsilon v_j\}}$$

- Sampling probabilities:

$$P_{S_j} = min\left\{1, \gamma b_{S_j}\right\}$$

γ is chosen such that $\sum_{i=1}^{q} P_{S_i} = \mathcal{B}$

Scheduling Probabilities and Mixing Matrix

- Nodes scheduling vector for k-th communication round

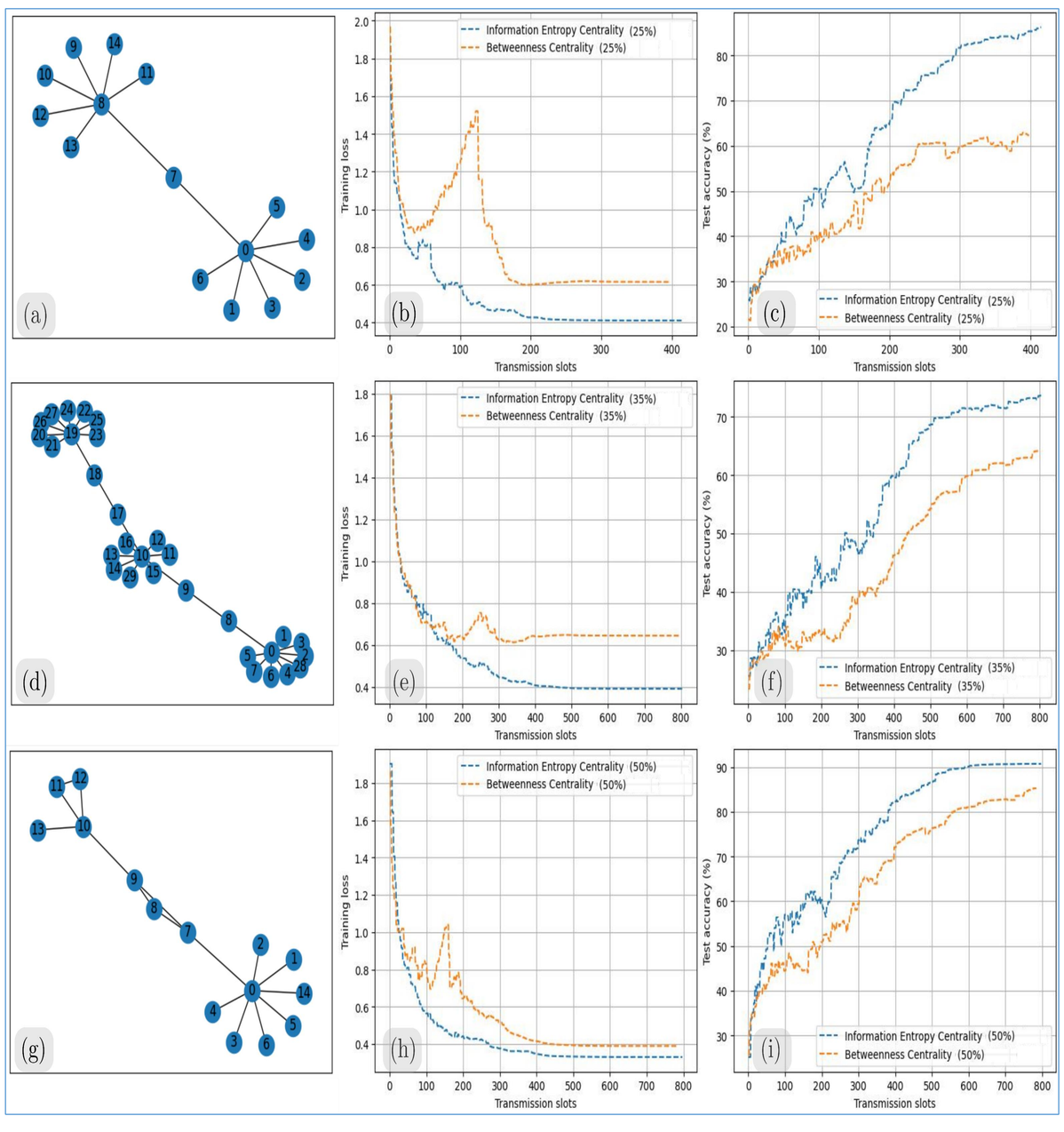
$$\mathbb{E}[Q(k)] = diag(p_1, \ldots, p_N) \qquad p_i = P_{S_j} \text{if } i \in V_r$$

- The subset probability had been optimized under constrained communication

$$\min_{P_{S_1},\ldots,P_{S_q}} \left\|\mathbb{E}\left(W^2(k) - \frac{1}{N}11^T\right)\right\|_2; \qquad \text{s.t.} \quad \sum_{l=1}^{q} P_{S_l} = \mathcal{B} \text{ and } 0 \le P_{S_l} \le 1, \forall l$$

- The solution of the above convex problem is

$$\min_{s,\alpha,\beta} s$$
$$\text{s.t. } \alpha^2 - \beta \le 0$$
$$I - 2\alpha\mathbb{E}[\tilde{L}^T(t)\tilde{L}(t)] + \beta\left(\mathbb{E}[\tilde{L}^T(t)\tilde{L}(t)]\right) - \frac{1}{N}11^T \preccurlyeq sI$$



Acknowledgments

The work of J. Nagar is supported by a Huawei France-EURECOM Chair on Future Wireless Networks. The work of P. A. Stavrou is supported in part by a Huawei France-EURECOM Chair on Future Wireless Networks and by the SNS JU project 6G-GOALS under the EU's Horizon programme Grant Agreement No. 101139232.