# Dimension Reduction Methods for Collaborative Mobile Gossip Learning

Árpád Berta, István Hegedűs
and Márk Jelasity

PDP16

# Motivation

- decentralized gossip learning in fully distributed networks (e.g. smart device network)
- dimension reduction is undiscovered area in this environment
- optimize communication cost:
  - message size → size of the learning model (e.g. weights) → number of the features
  - in case of high cost the learning is infeasible in extreme scenario

# Dimension reduction

- reduce number of the features from the raw data
  - e.g.: raw data size of text, image, or activity recognition data
- feature extraction:
  - **d** size feature-space project to its **k** size subspace
  - examined methods (linear projections):
    - Singular Value Decomposition (SVD)
    - Random Projection selection (RP)
    - hybrid of both (SVDRP)

# System Model

- potentially large number of nodes (personal computers, smart sensors, wearable devices, phones, tablets, etc.) with only one training data
- communicate with their neighbors via messaging
- neighbors: peer sampling service
- nodes can leave the network and join again without any prior notice (with unchanged state)

# Gossip Learning

- learning models perform random walk in the network:
  - every node update the received model with local training data:
    - stochastic gradient descent (SGD) step
  - and send it toward immediately (hot potato)
- privacy preserving: local data never leaves the node

# Singular Value Decomposition

- István Hegedűs, Márk Jelasity, Levente Kocsis, A. András Benczúr - Fully distributed robust singular value decomposition. (P2P2014)
- SGD SVD
- Message size: the whole projection matrix $\rightarrow O(k \cdot d)$
- Once we did this with expensive communication cost then the reduced data makes possibility of cheap communication cost in different learning task

# Random projection selection

- cheap to generate random projection matrix (sparse) in every node
- learning model on previously reduced data
- approximate the error with flying average on training examples

→ possibility of chose the best projection matrix

# Random projection selection

- maximize the learning accuracy on the reduced data → **two models** in on message:
  - a model based on current examined projection matrix
  - the best projection based on the approximated error
- every node process the same random matrix based on the same **seed**
- Message size: $2 \cdot (k+C)$

# SVD-RP Hybrid



- quickly RP: small messages, fast converged but moderate accuracy
- meanwhile SVD is trying to converge, vector orthogonality:

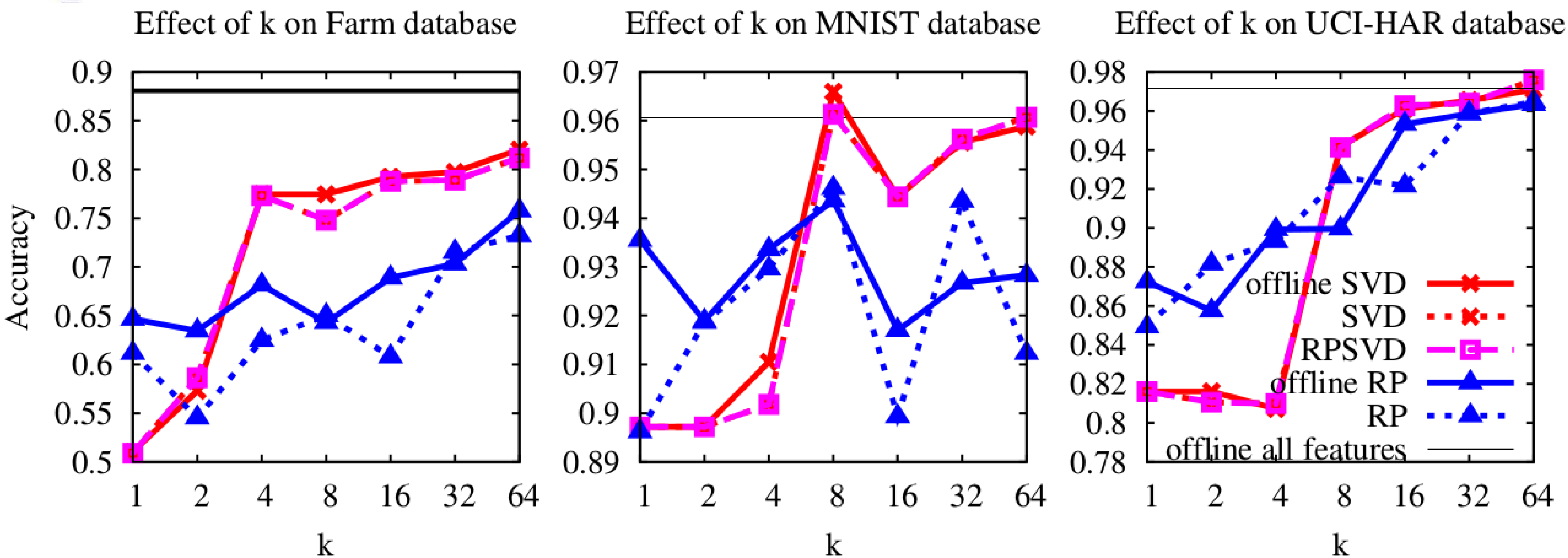
$$o_i = \frac{1}{k-1} \sum_{j \neq i} \frac{p_i^T \cdot p_j^T}{\|p_i^T\| \|p_j^T\|},$$

- as it happens, change to SVD: huge messages but good accuracy

# Experimental setups

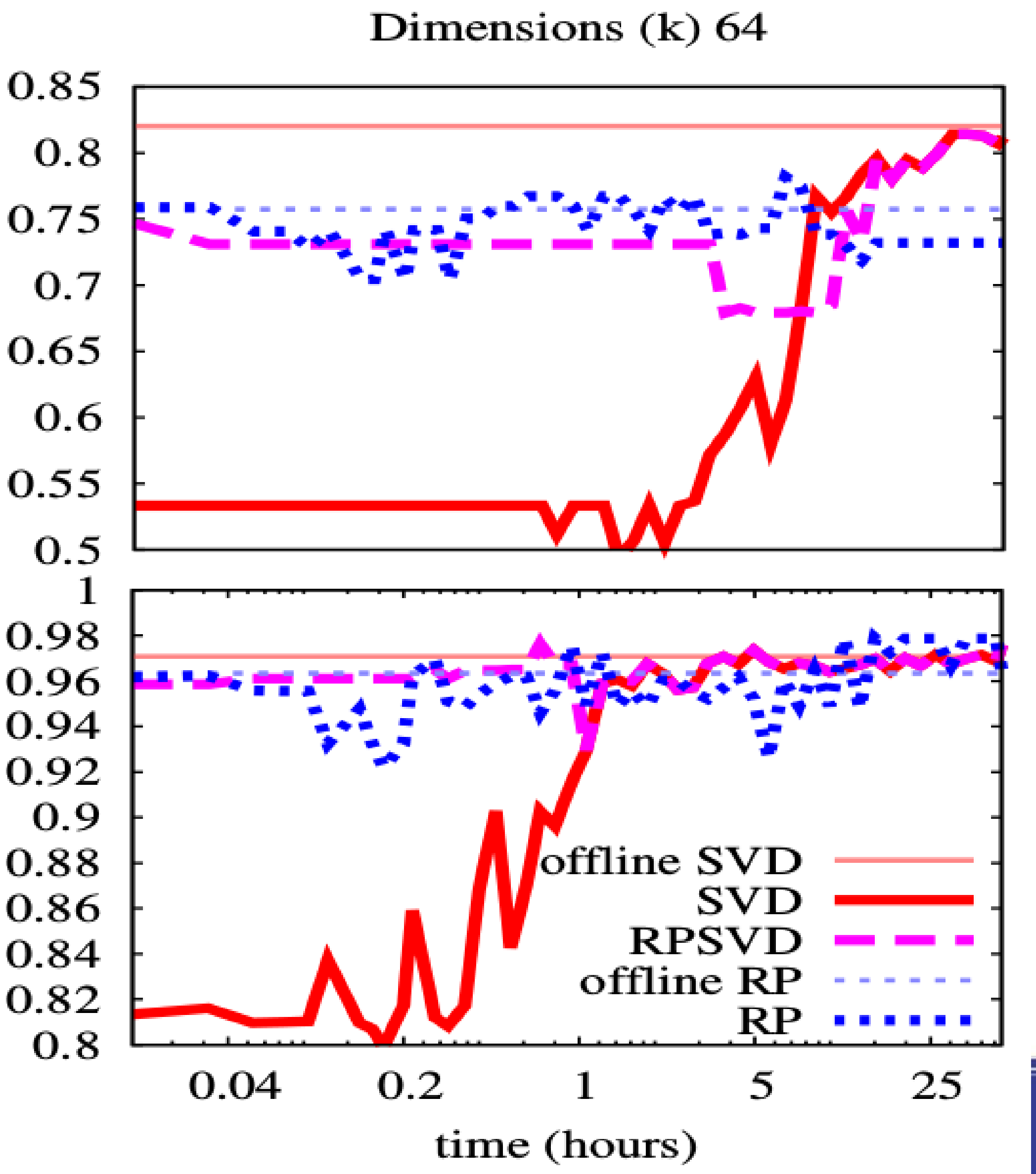
- **real datasets** (text -, image processing, activity recognition) – evaluate on test data
- network size is based on the number of the training data (every node has only one example)
- static network, 50 neighbor
- **real churn** – smart phone trace
- **realistic communication costs**, nodes initially start
  - SVD: 100% SVD model
  - RP: 100% RP model
  - SVDRP: 99% SVD, 1% RP model
- evaluated with **Logistic Regression** learning accuracy

# Results




| | MNIST | Farm ads | HAR |
|---|---|---|---|
| Training set size | 60 000 | 3 314 | 7 352 |
| Test set size | 10 000 | 829 | 2947 |
| Number of features | 784 | 54 877 | 561 |
| Original number of classes | 10 | 2 | 6 |
| Positive examples | 10% | 53% | 17% |

# Results (k=64)



Dimensions (k) 64
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
offline SVD
SVD
RPSVD
offline RP
RP
0.04
0.2
1
5
25
time (hours)