UNIVERSITAS SCIENTIARUM SZEGEDIENSIS

Research Group on Artificial Intelligence

Gossip-Based Machine Learning in Fully Distributed Environments

István Hegedűs

Márk Jelasity
supervisor

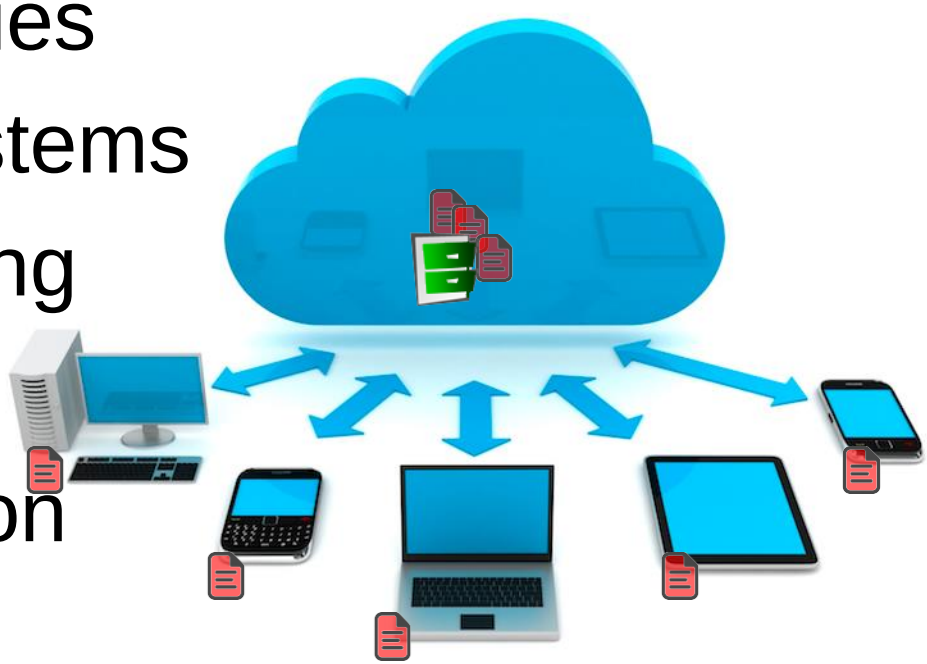
University of Szeged
MTA-SZTE Research Group on AI
Hungary

Motivation

- Data is accumulated in data centers
- Costly storage and processing
  - Maintenence, Infrastructure, Privacy
- Limited access
  - For researchers as well
- But, data was produced by us

Motivation – ML Applications

- Personalized Queries
- Recommender Systems
- Document Clustering
- Spam Filtering
- Image Segmentation

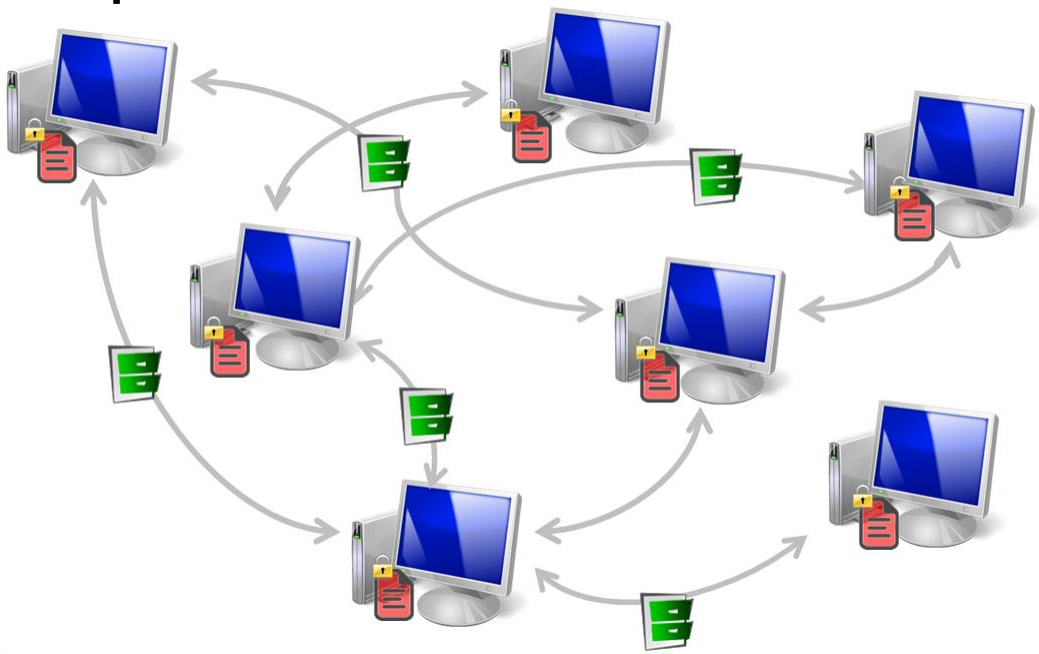


Gossip Learning

- ML is often an optimization problem
- Local data is not enough

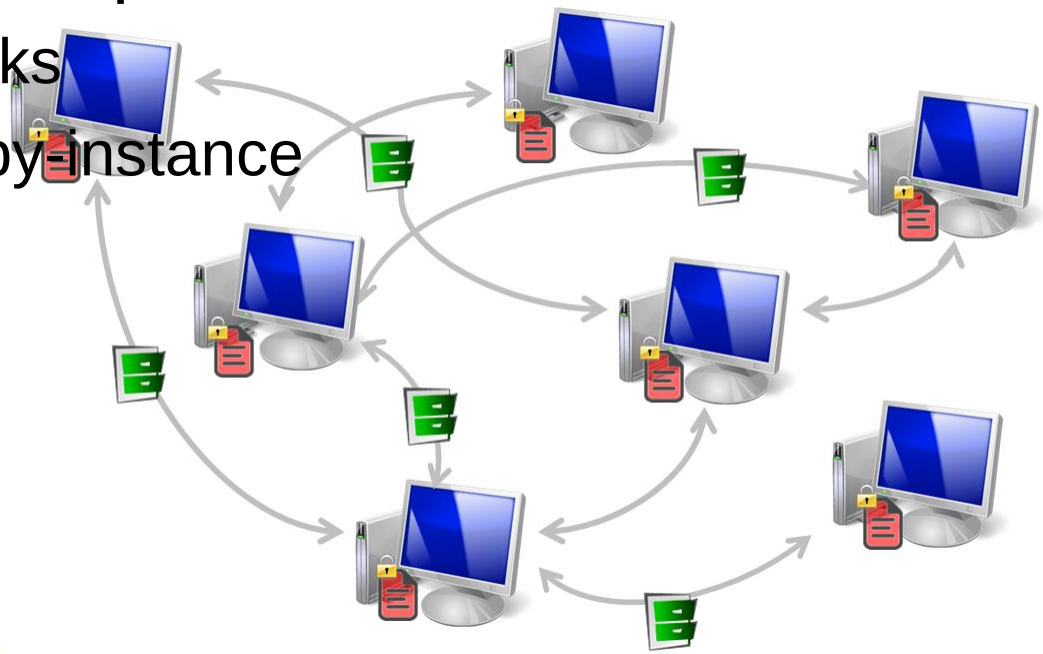
Gossip Learning

- ML is often an optimization problem
- Local data is not enough
- Models are sent and updated on nodes



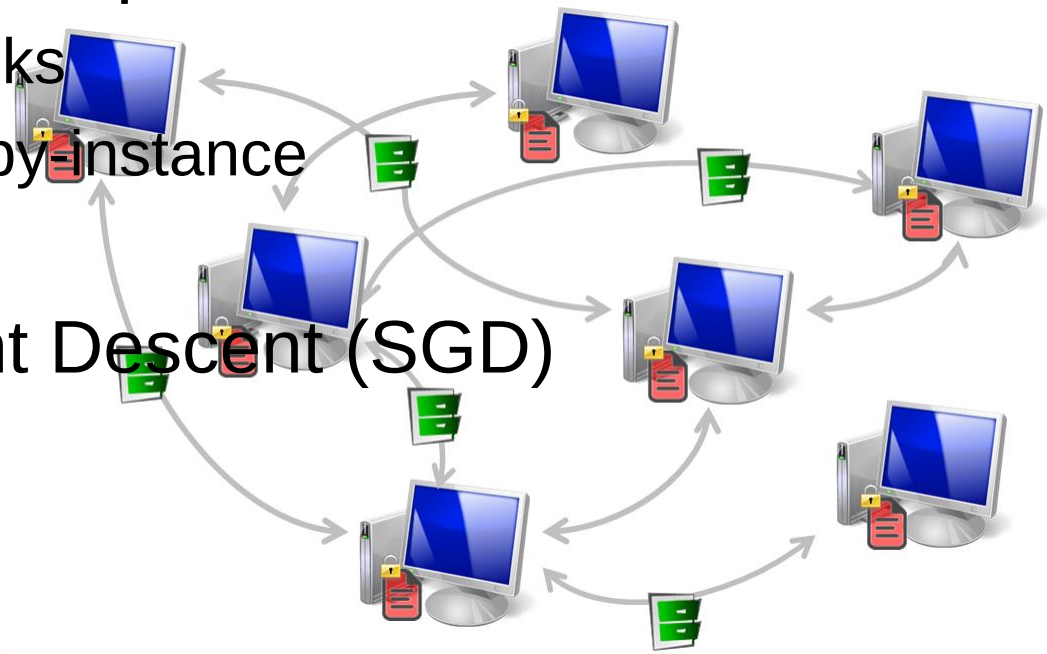
Gossip Learning

- ML is often an optimization problem
- Local data is not enough
- Models are sent and updated on nodes
  - Taking random walks
  - Updated instance-by-instance
  - Data is never sent



Gossip Learning

- ML is often an optimization problem
- Local data is not enough
- Models are sent and updated on nodes
  - Taking random walks
  - Updated instance-by-instance
  - Data is never sent
- Stochastic Gradient Descent (SGD)



SGD

- Objective function

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

SGD

- Objective function

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

$$w_{t+1} = w_t - \eta_t\left(\frac{\partial J}{\partial w}\right)$$

$$= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right)$$

SGD

- Objective function

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

$$w_{t+1} = w_t - \eta_t\left(\frac{\partial J}{\partial w}\right)$$

$$= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right)$$

- SGD, data can be processed online (instance by instance)

$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$

SGD

- Objective function

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

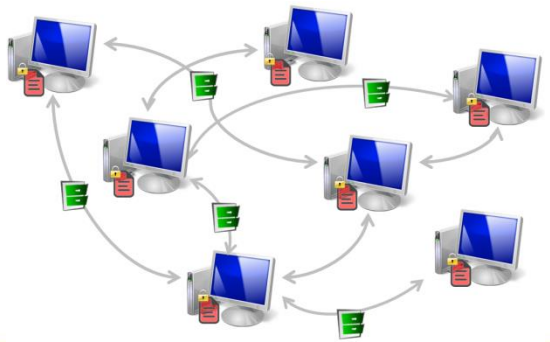
$$w_{t+1} = w_t - \eta_t\left(\frac{\partial J}{\partial w}\right)$$

$$= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right)$$

- SGD, data can be processed online (instance by instance)

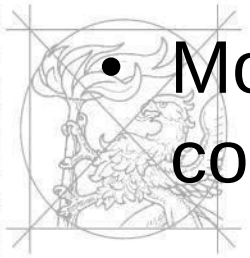
$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$

- Gossip Learning



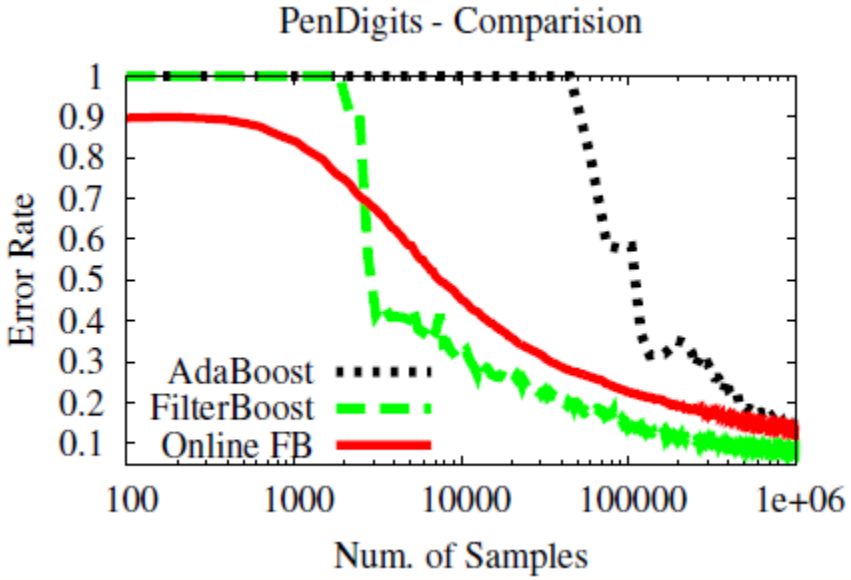
Gossip-Based Learning

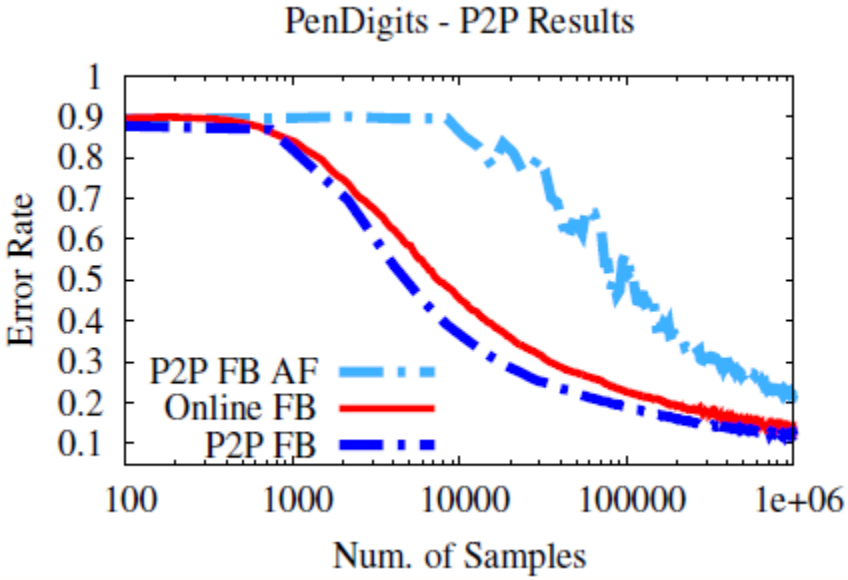
- SGD-based machine learning algorithms can be applied, e.g.
  - Logistic Regression
  - Support Vector Machines
  - Perceptron
  - Artificial Neural Networks
- Training data never leave the nodes
- Models can be used locally additional communication is not required



Boosting

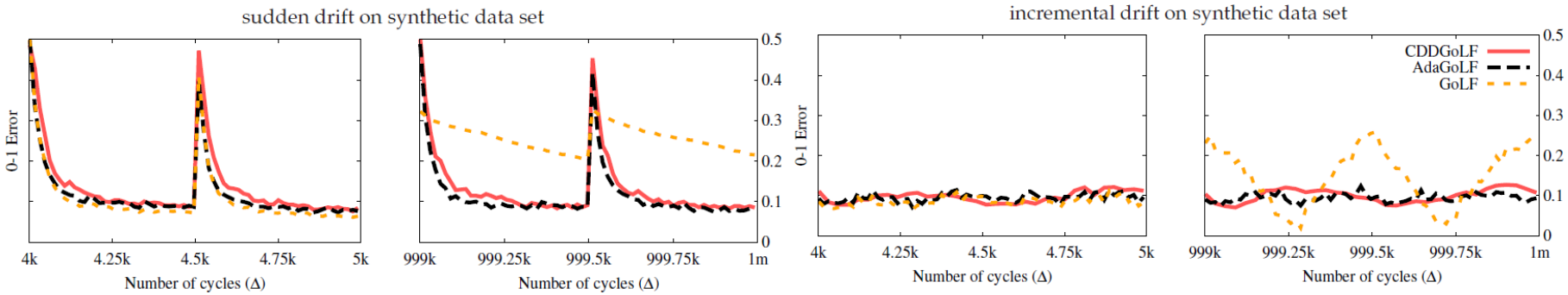
- Boosting is achieved by online weak learning
- Online FilterBoost is proposed
- Results are competitive to AdaBoost method





Handling Concept Drift

- Two adaptive learning mechanisms by
  - Managing model age distribution
  - Model performance monitoring
- Drift handling and detection capabilities

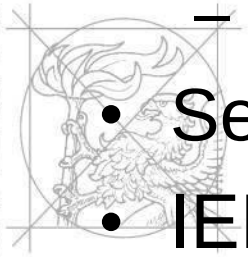


SVD

- SGD based low-rank matrix approximation

$$J(X,Y) = \frac{1}{2}\|A - XY^T\|_F^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij} - \sum_{l=1}^{k} x_{il}y_{jl})^2$$

- A modification that converges to the SVD
- Can be used for
  - Recommender systems
  - Dimension reduction
- Sensitive data never leave the nodes as well
- IEEE P2P'14 best paper



Conclusion

- A possible way of machine learning on fully distributed data was proposed
- A gossip-based framework was presented with numerous learning algorithms
  - Logistic regression, SVM, Perceptron, Boosting, SVD
- Concept drift handling capabilities were improved as well

Related Publications

	Chapter 3	Chapter 4	Chapter 5	Chapter 6
CCPE 2013 [6]	●	○	○	○
EUROPAR 2012 [3]	○	●		
SASO 2012 [4]	○		●	
SISY 2012 [2]	○		●	
ACS 2013 [5]	○		●	
P2P 2014 [9]	○			●
EUROPAR 2011 [1]	○			
ICML 2013 [7]	○			
ESANN 2014 [8]	○			
TIST 2016 [11]	○			○
PDP 2016 [12]	○			
PDP 2016 [10]	○			○

Questions(Alberto Montresor)

What are the advantages of executing your approach not in completely decentralized systems (like P2P networks), but instead in a cluster of distributed machines. This should be answered for all the proposed techniques.

Questions (Attila Kiss) I.

In these algorithms, nodes exchange model parameters. While this is better than sharing personal data, it is well-known that exchanging such information can still leak some sensitive information about the data used to compute these parameters/gradients. In machine learning, the most popular notion of privacy is differential privacy, which gives strong probabilistic guarantees. Differential privacy can be achieved by adding noise to various quantities: either the data itself, the model updates, the objective function, or the output (see e.g. C. Dwork. Differential privacy: A survey of results. In Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, pages 1-19, 2008.)Could the algorithms in the thesis be extended merits and drawbacks in terms of convergence rate and communication cost?

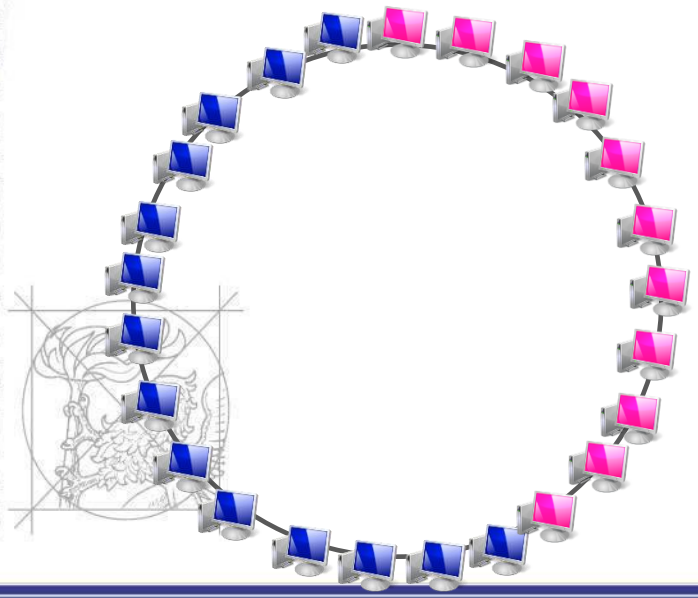
Questions (Attila Kiss) II.

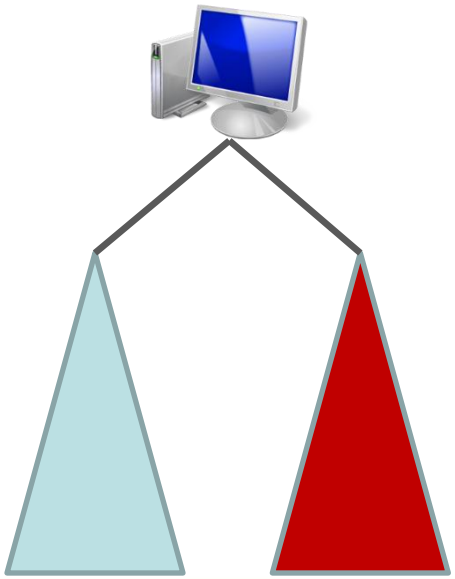
The author assumes that the homogenous network graph reflects the similarity between nodes (i.e., neighbors in the network graph have similar objectives). However, in practical scenarios, nodes could be different, one node can store larger or more reliable data than the other nodes, communicates faster, has more computing capacity or providing more useful information. This requires strategies to discover good peers and combining this information with the algorithms in the thesis to obtain more efficient decentralized protocols. What could be a good trade-off between exploration and exploitation in peer discovery to improve decentralized learning?

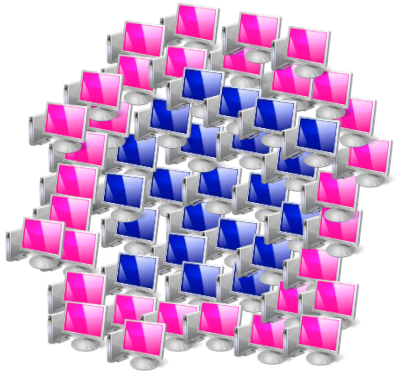
Questions (Attila Kiss) III.

What is the impact of the network topology on the convergence speed of the algorithm in the thesis? How does this speed depend from the usual graph parameters e.g. from clustering coefficient of the network in general or in special cases?

Topológia függő adateloszlások







Questions (Attila Kiss) IV.

Could the author give negative cases, machine learning methods in the field of classification, clustering or association rules, where gossip based approach is not applicable?