Networked Federated Learning

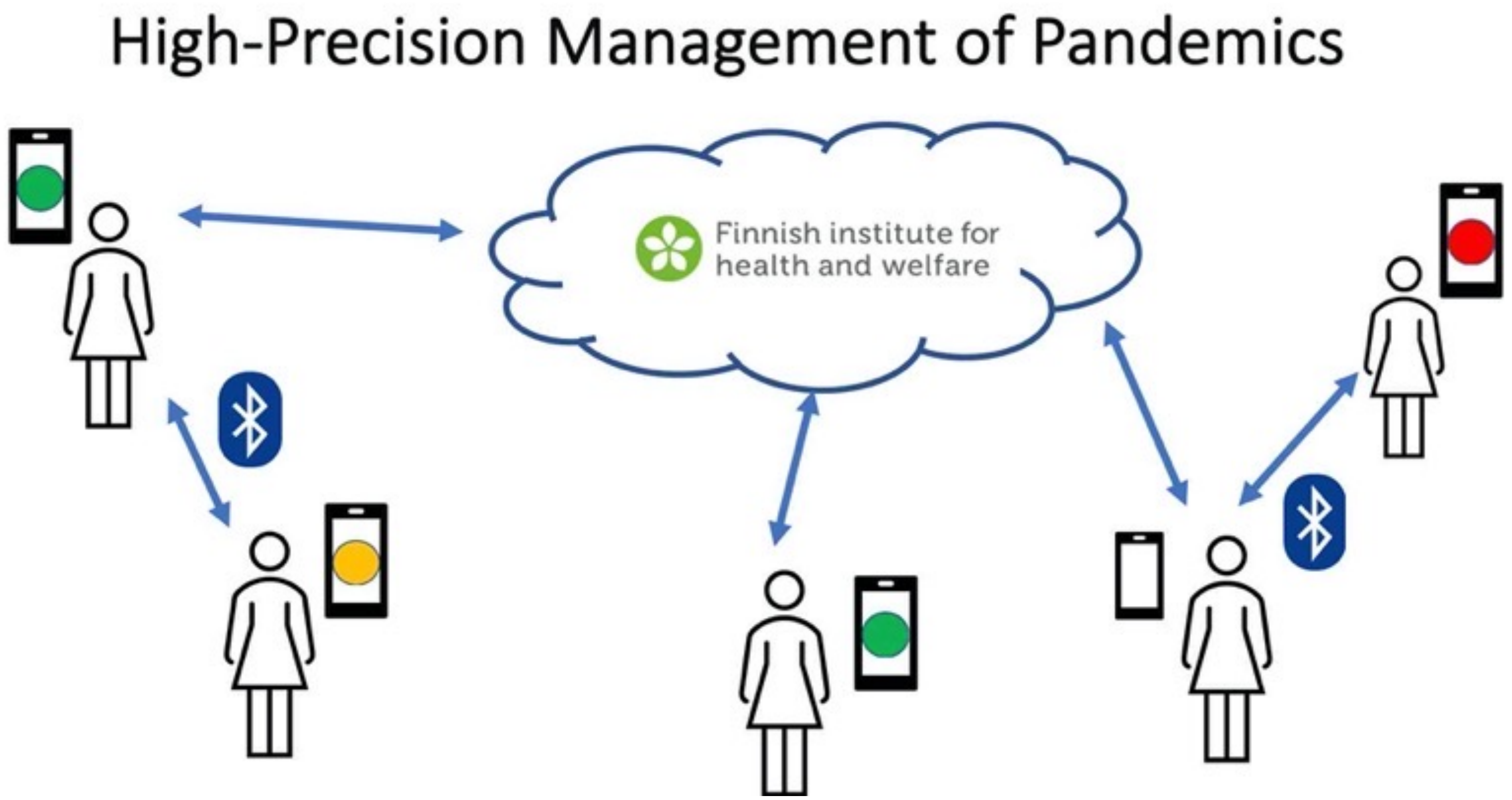
Alexander Jung, Feb. 2022

https://www.linkedin.com/in/aljung/
https://www.youtube.com/channel/UC_tW4Z_GfJ2WCnKDtwMuDUA
https://twitter.com/alexjungaalto

About Me.

- MSc (2008) and Ph.D. (2012) in EE, TU Vienna
- since 2015 Ass. Prof. for Machine Learning at Aalto/CS
- leading group “Machine Learning for Big Data”
- two current main research areas (RA)
- teaching ML courses at Aalto and fitech.io

RA1: Networked Federated Learning.

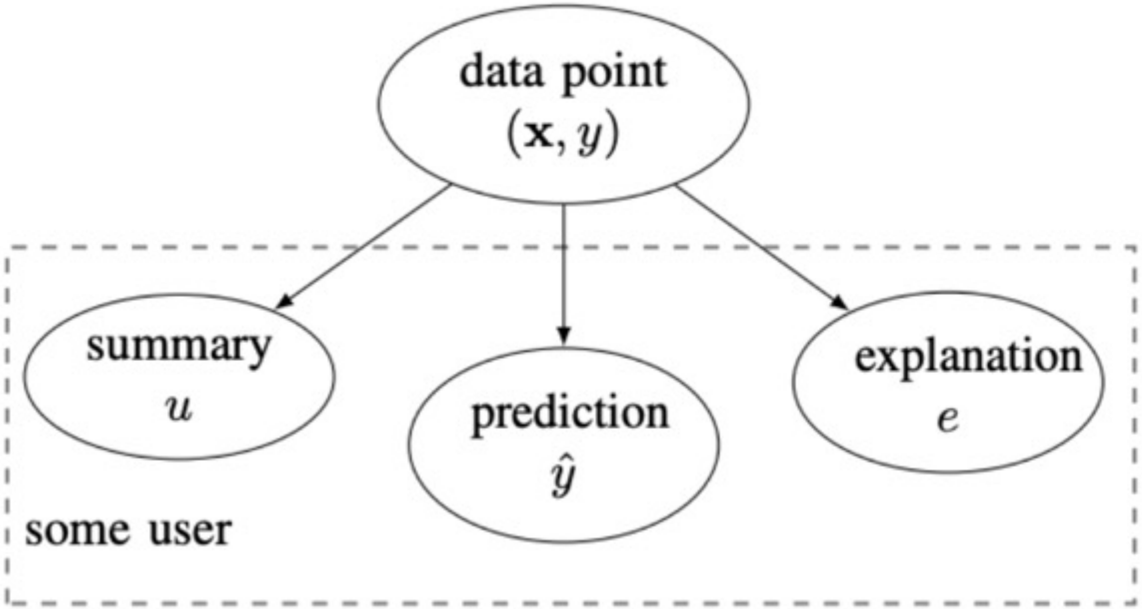


Y. Sarcheshmehpour, M Leinonen and AJ, “Federated Learning From Big Data Over Networks”, IEEE ICASSP, 2021.

AJ, “Networked Exponential Families for Big Data Over Networks,” in IEEE Access, 2020, doi: 10.1109/ACCESS.2020.3033817.

AJ, N. Tran, “Localized Linear Regression in Networked Data,” in IEEE SPL, 2019, doi: 10.1109/LSP.2019.2918933.

RA2: Explainable Machine Learning.



explanation can be:

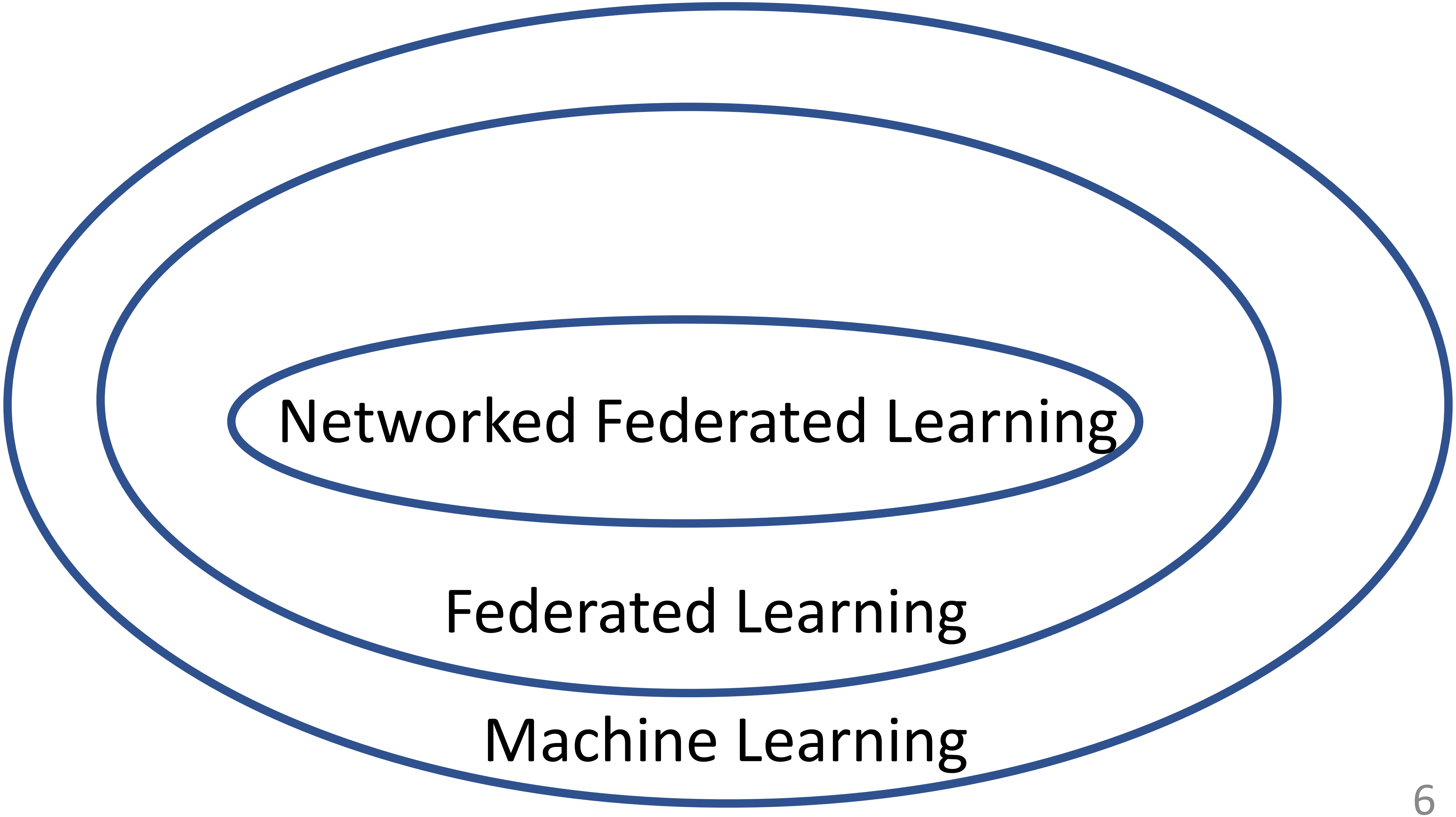
- relevant example of training set
- subset of features
- counterfactuals
- a free text explanation
- court sentence

AJ, “Explainable Empirical Risk Minimization”, arXiv eprint, 2020. [weblink]

AJ Jung and P. H. J. Nardelli, “An Information-Theoretic Approach to Personalized Explainable Machine Learning,” in IEEE SPL, 2020, doi: 10.1109/LSP.2020.2993176.

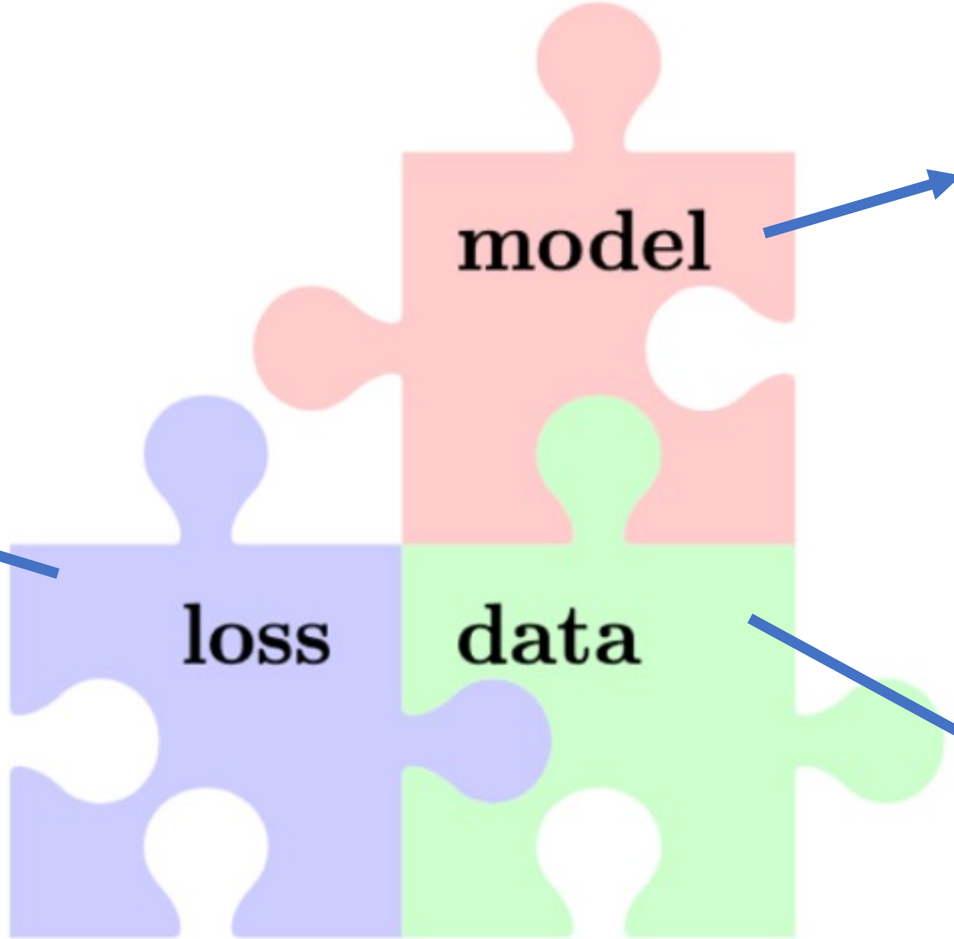
Networked Federated Learning

In a nutshell:
organize data, models and computation for machine learning as networks.


Networked Federated Learning
Federated Learning
Machine Learning

Three Components of ML

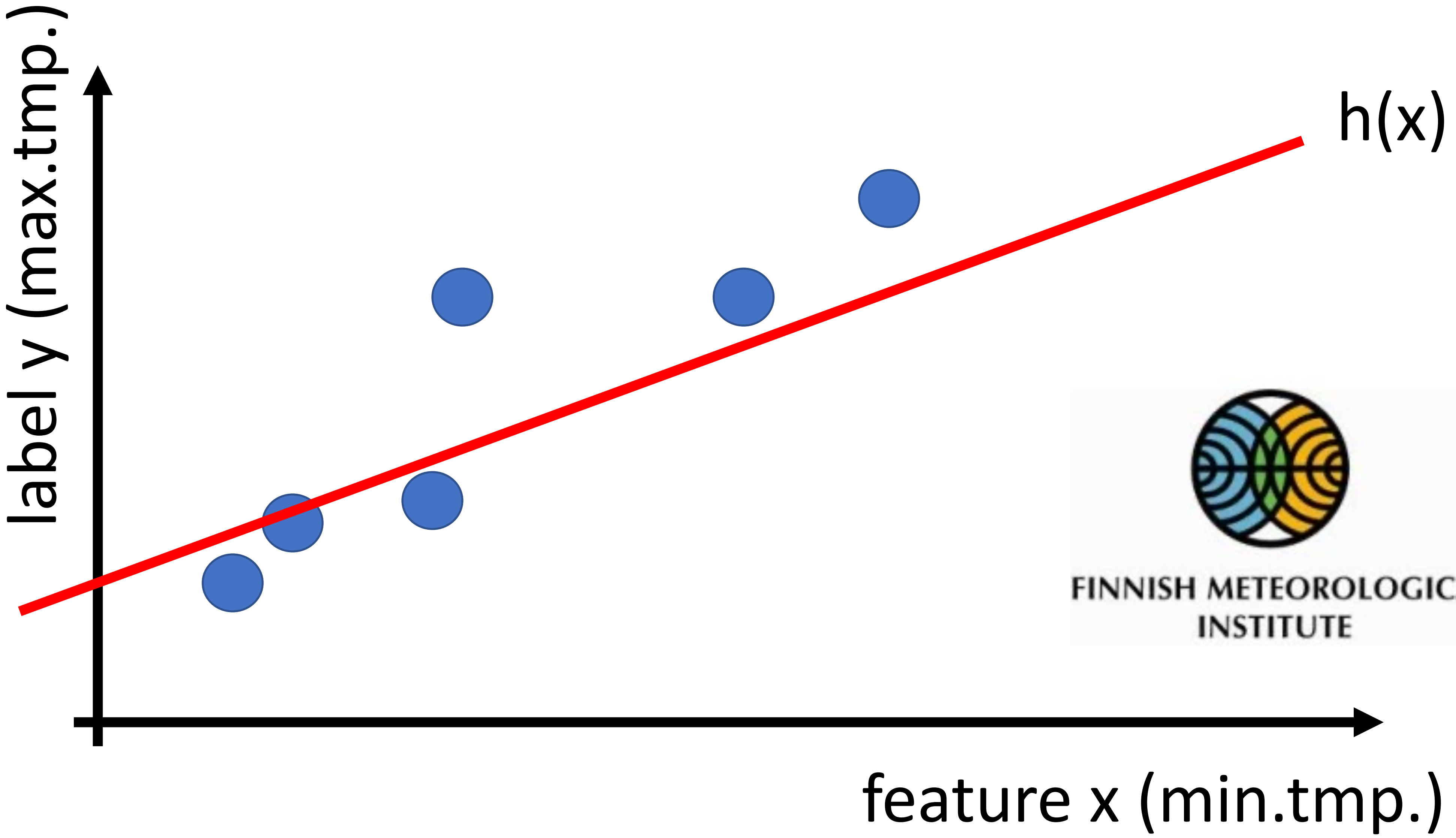
- squared error
- logistic loss
- focal loss
- ….

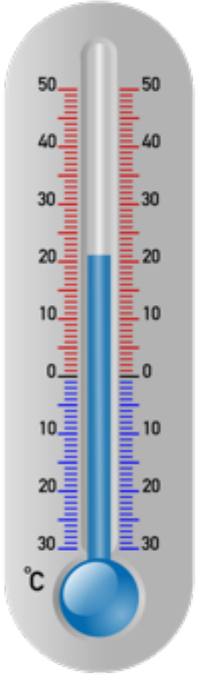


- linear space
- decision tree
- neural nets
- ….

- numeric arrays
- text
- networks
- …

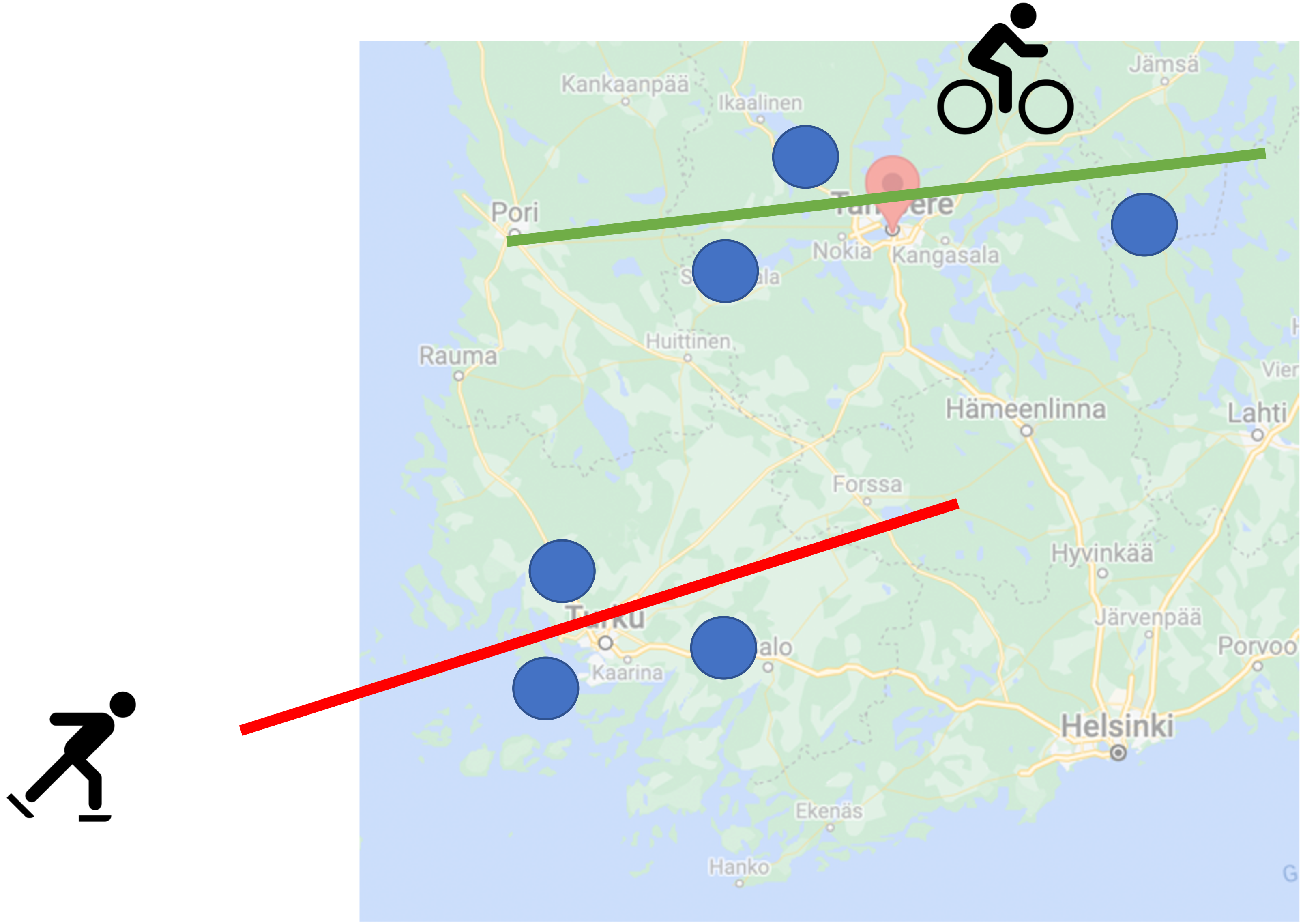
Plain Old Machine Learning.





FINNISH METEOROLOGICAL INSTITUTE

Networked Federated Learning



Networked Data

Weather Stations.

FINNISH METEOROLOGICAL INSTITUTE

ImageNet.

"…ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images…"

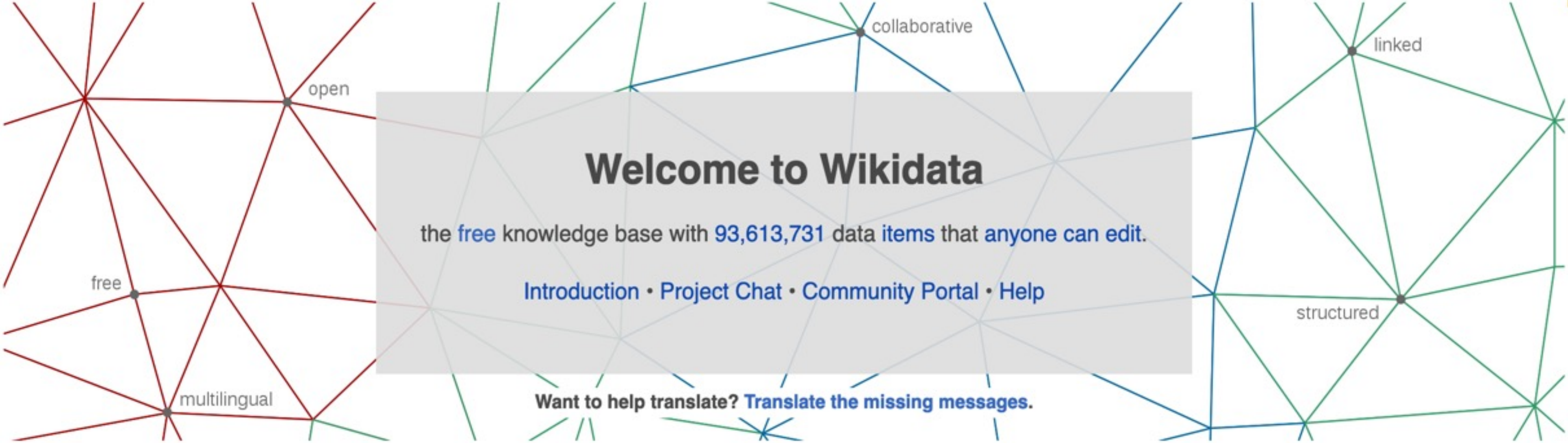
https://image-net.org/

WordNet.

"…Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept… The resulting network of meaningfully related words and concepts can be navigated….."

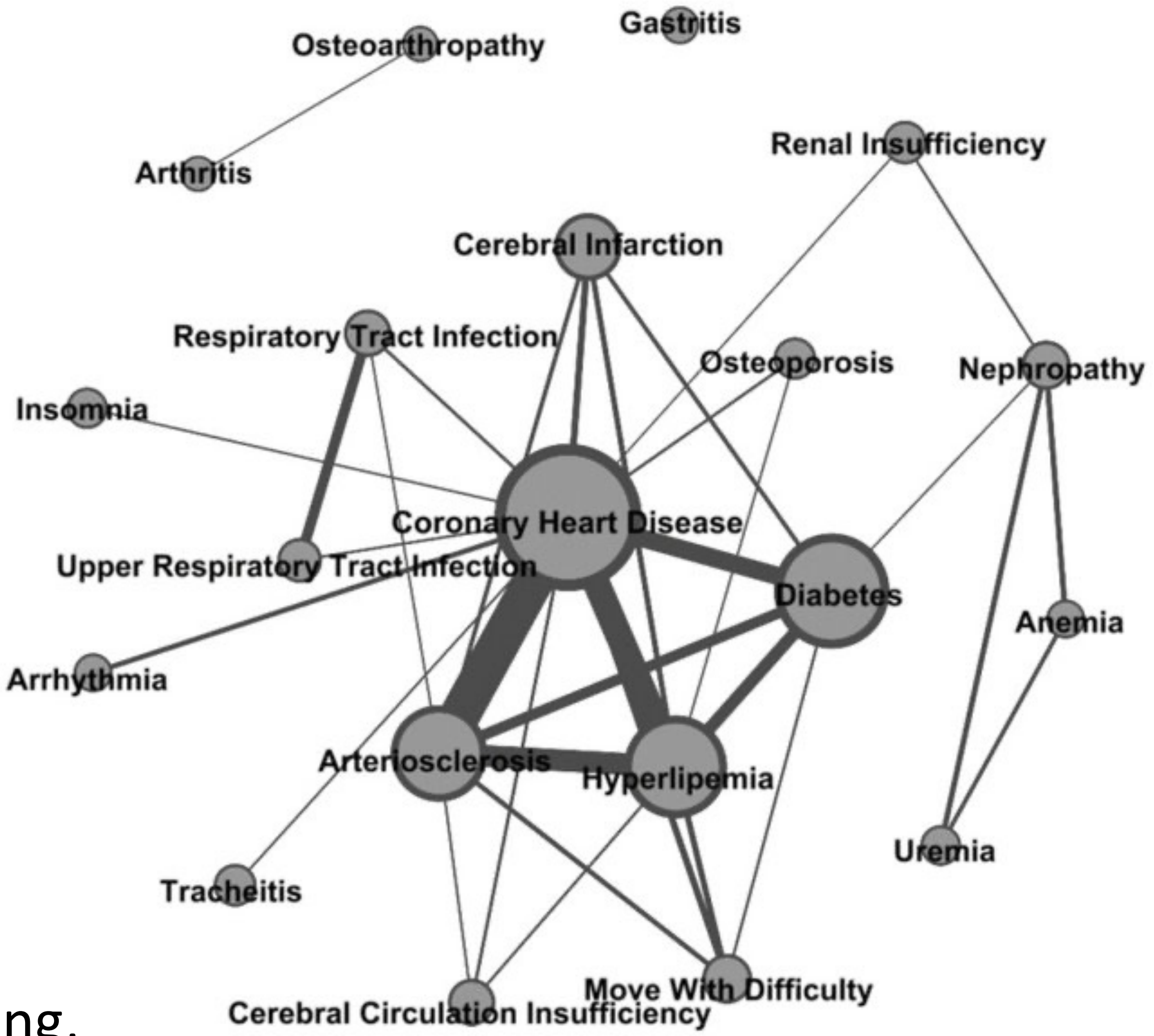
https://wordnet.princeton.edu/

Wikidata.



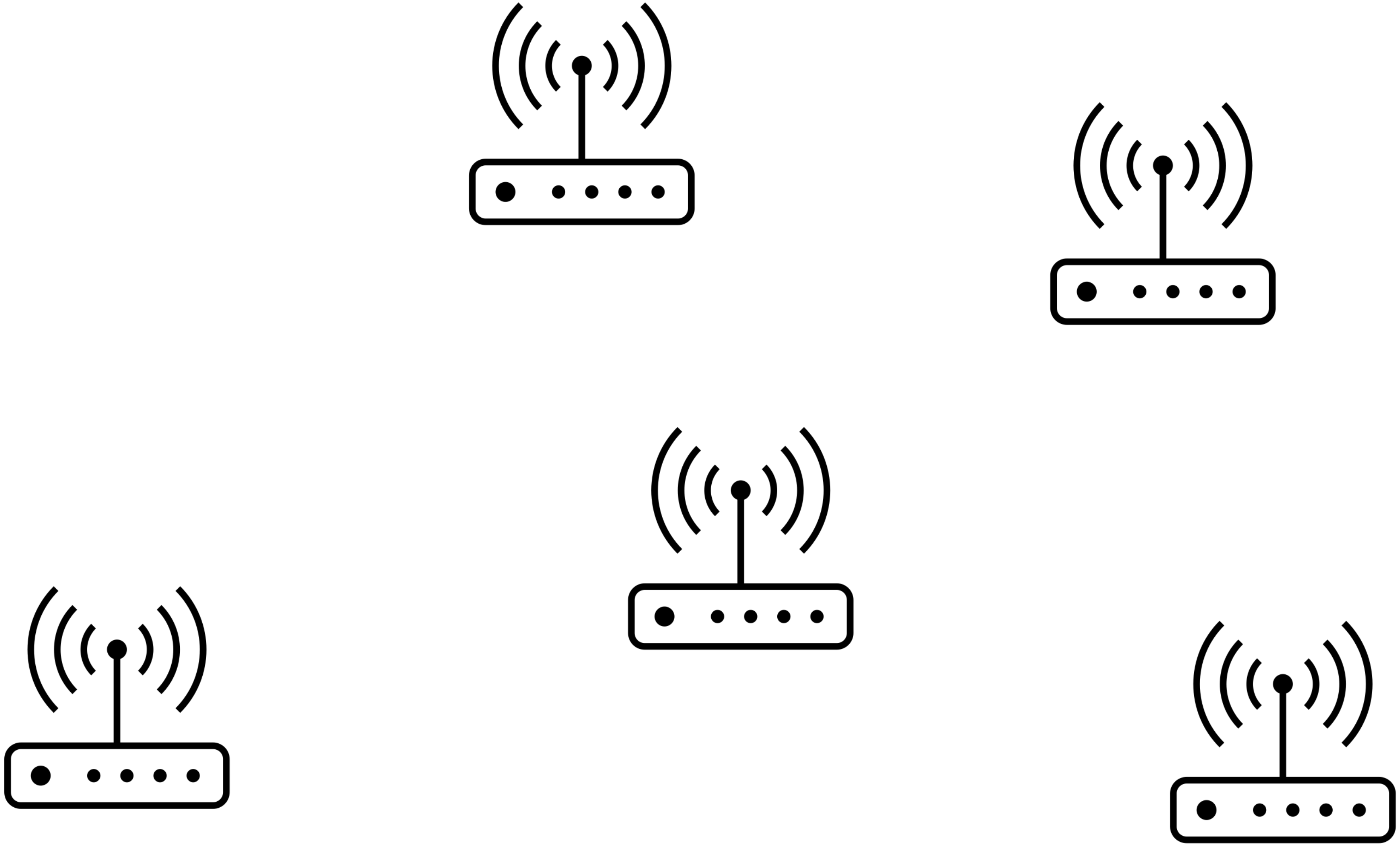
https://www.wikidata.org/wiki/Wikidata:Main_Page

Diseases.



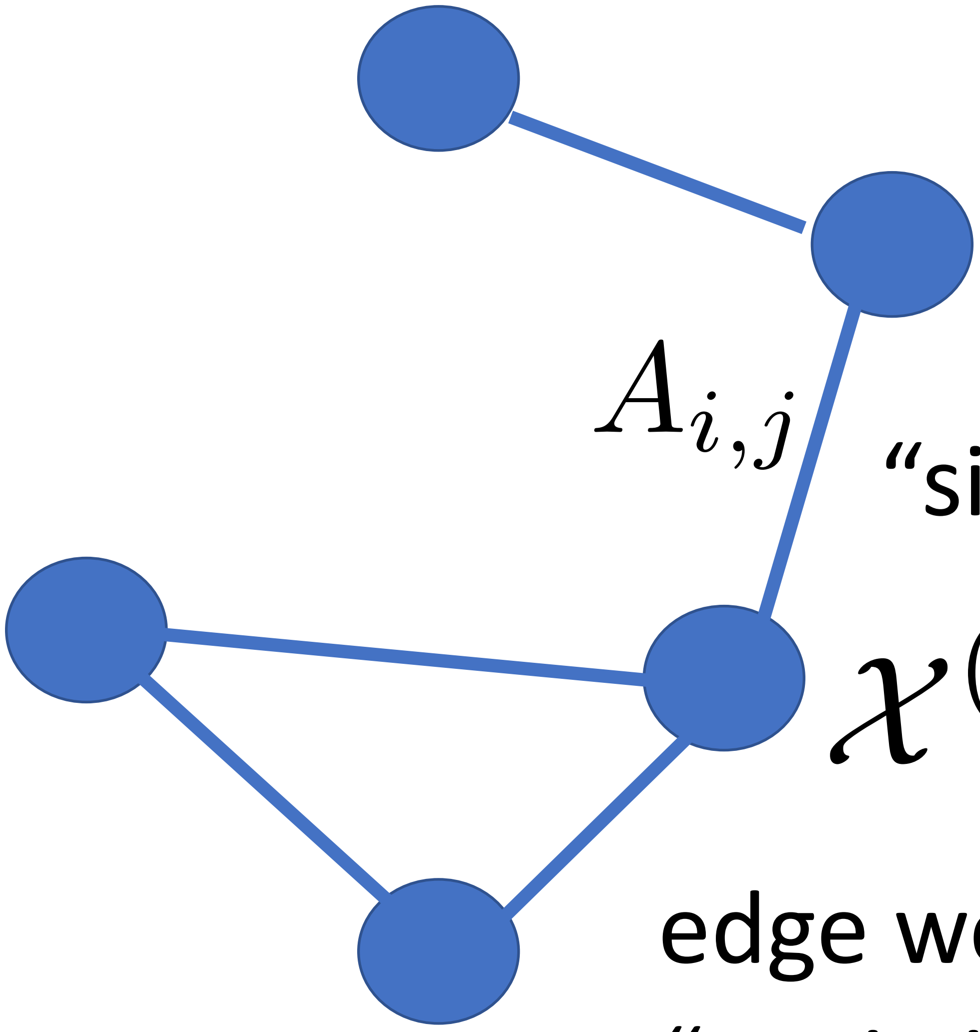
Liu, Jiaqi & Ma, James & Wang, Jiaojiao & Zeng,
Daniel Dajun & Song, Hongbin & Wang, Ligui & Cao, Zhidong. (2016).
Comorbidity Analysis According to Sex and Age in Hypertension Patients in China.
International Journal of Medical Sciences. 13. 99-107. 10.7150/ijms.13456.

WSN.



Anchors.

Abstraction – The Empirical Graph.



local dataset $\mathcal{X}^{(i)}$

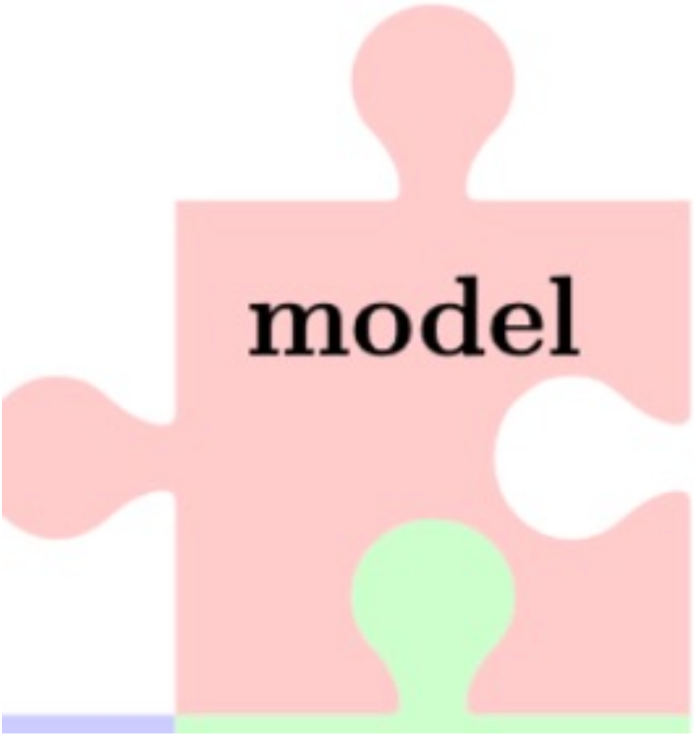
$A_{i,j}$

“similar"

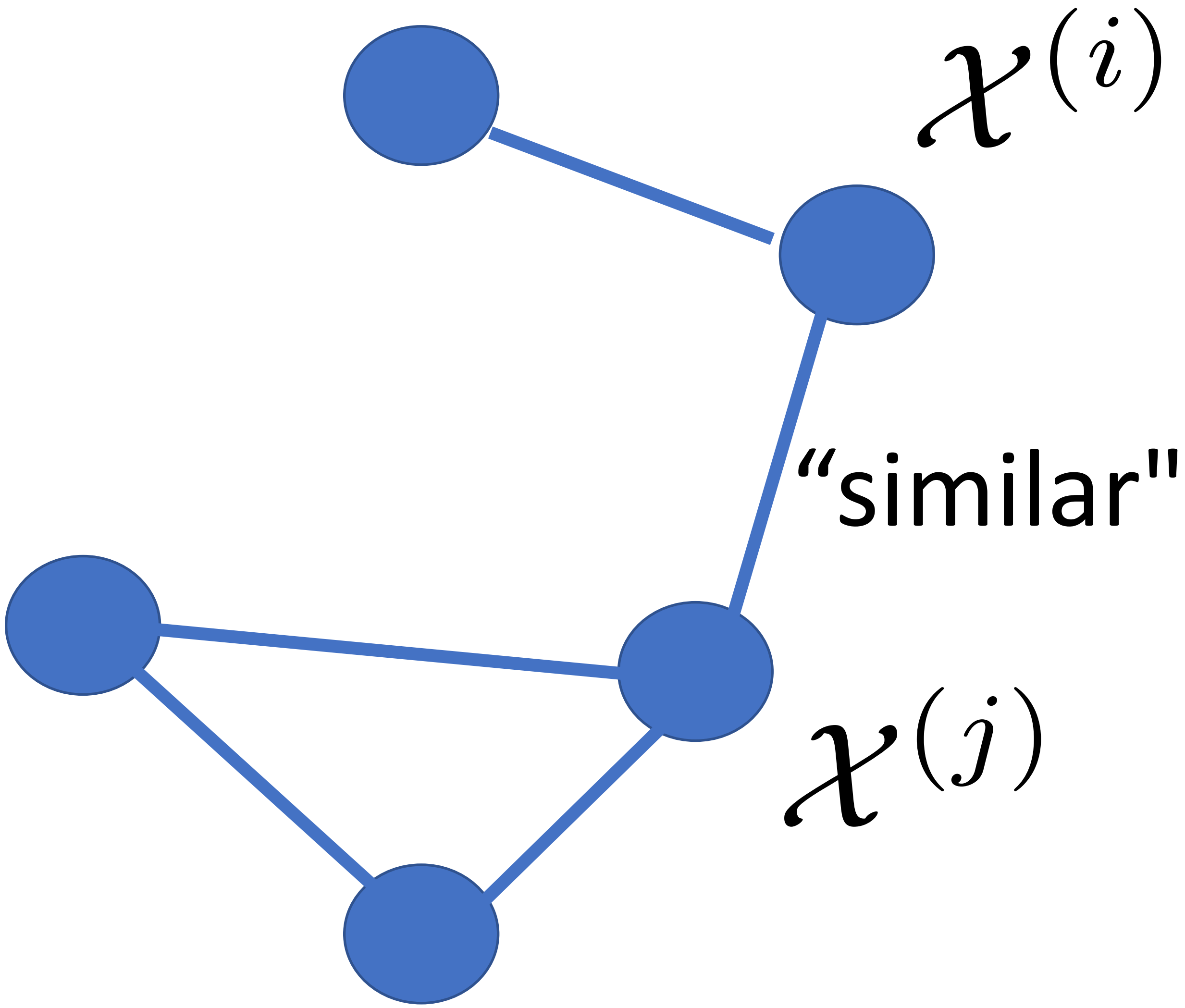
$\mathcal{X}^{(j)}$

edge weights $A_{i,j}$ quantify
“statistical similarities”

Networked Models



Networked Models.

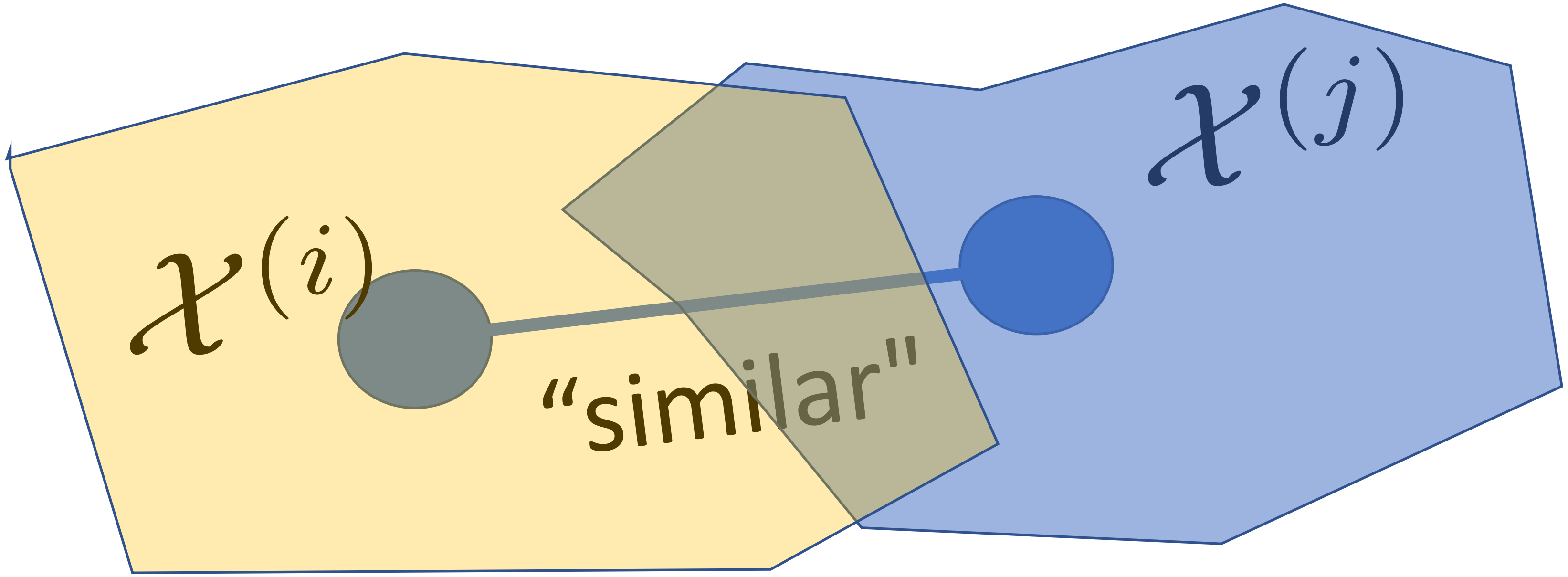


local model for each node

couple models at connected nodes

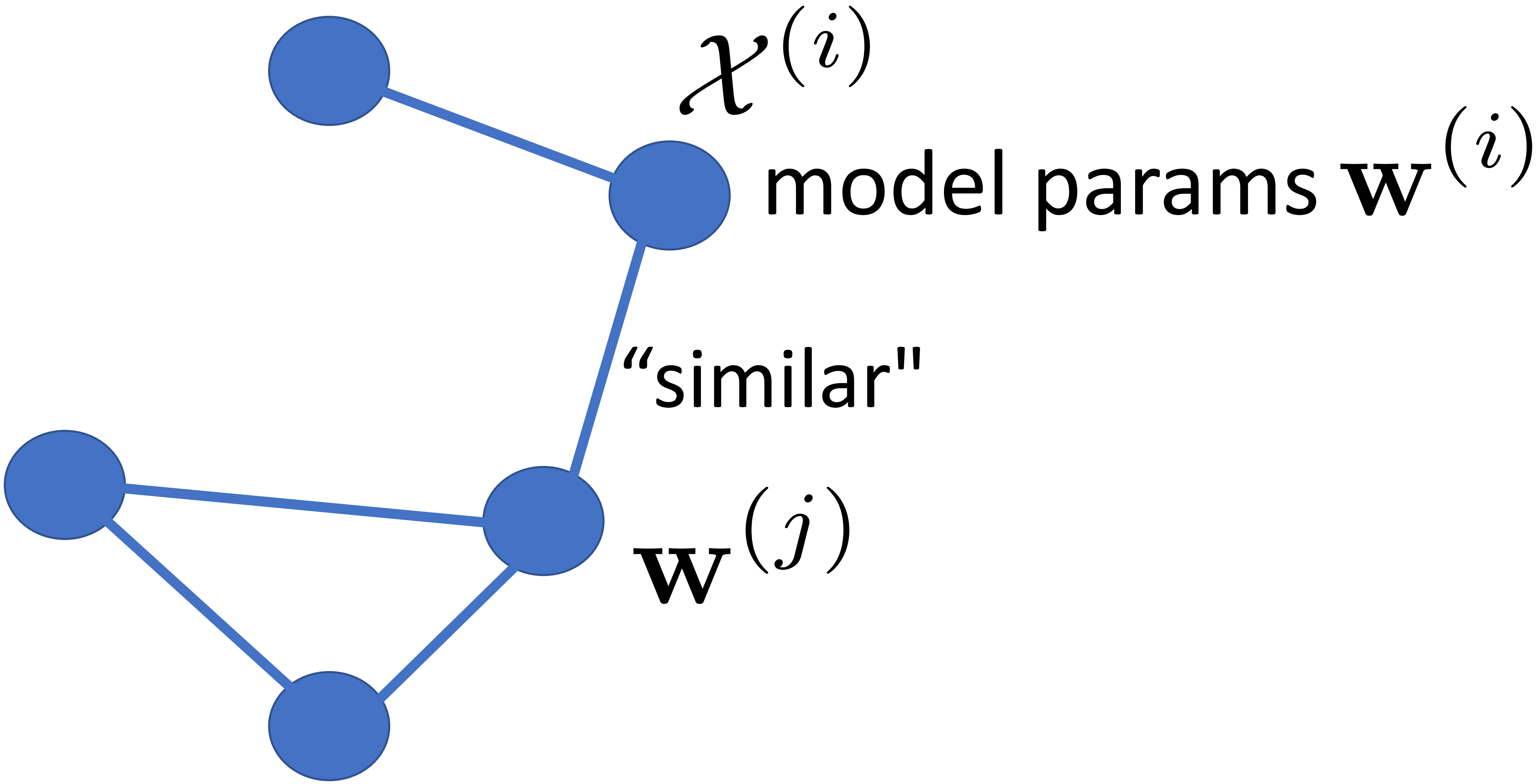
Sheaves on Graphs.

Definition 2.2 (Sheaves). Let G be a graph. A sheaf $\mathcal{F}$ on G consists of a vector space $\mathcal{F}(v)$ for each vertex v of G, a vector space $\mathcal{F}(e)$ for each edge e of G, and a linear transformation $\mathcal{F}_{v \trianglelefteq e} : \mathcal{F}(v) \to \mathcal{F}(e)$ for each incident vertex-edge pair $v \trianglelefteq e$.

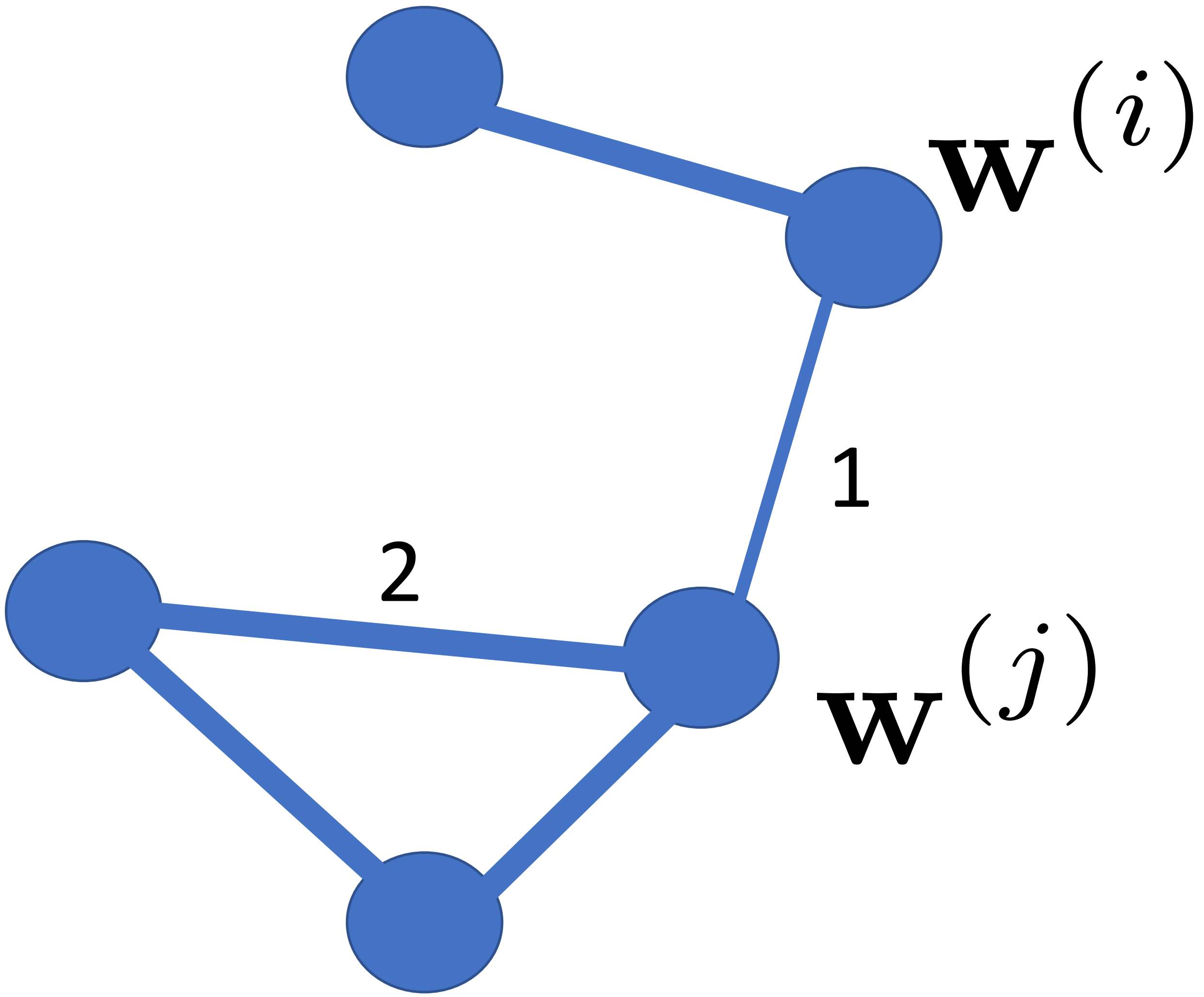


https://www.jakobhansen.org/publications/gentleintroduction.pdf

Networked Parametric Models.



Generalized Total Variation (GTV)



force params of well connected nodes to be similar by requiring a small GTV

$$\sum_{\{i,j\}} A_{i,j} \phi\big(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\big)$$

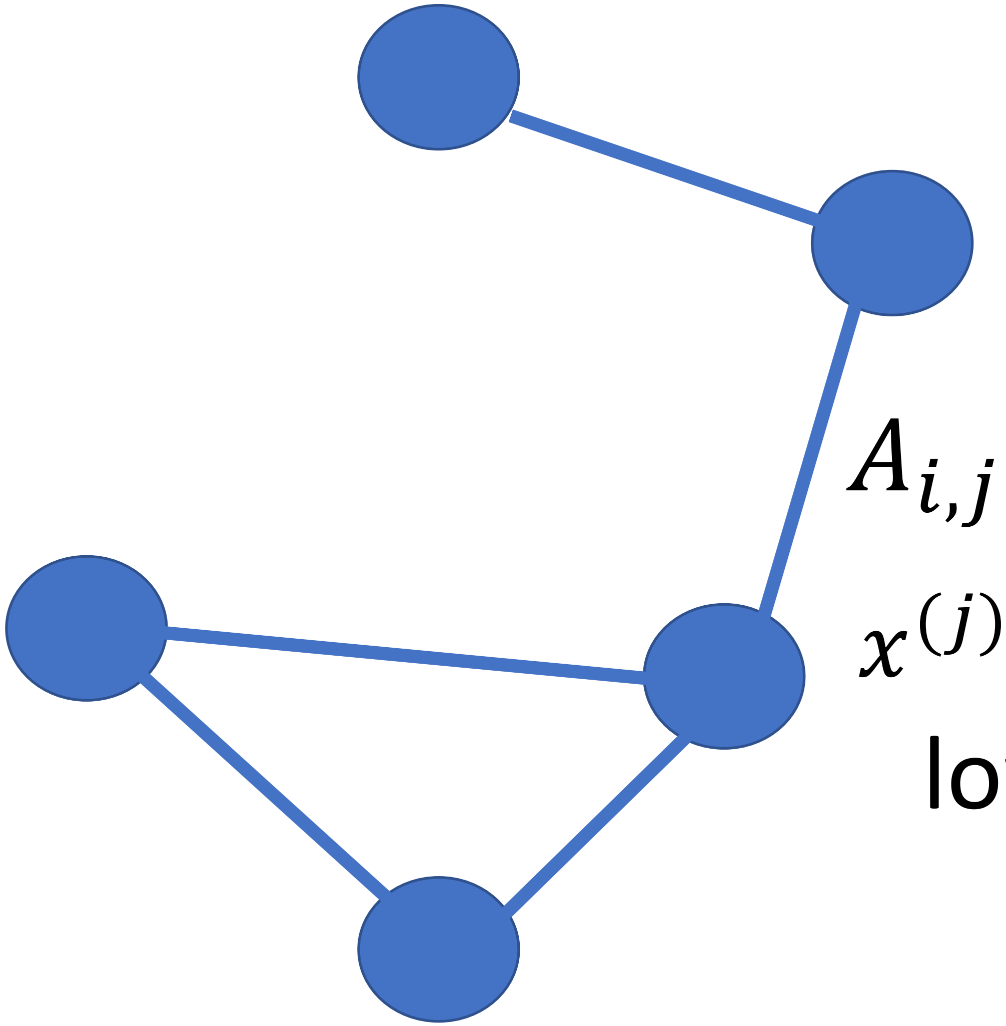
Two Special Cases of GTV.

total variation $\phi(\mathbf{u}) = \|\mathbf{u}\|_2$

graph Laplacian quadratic from is GTV with

$$\phi(\mathbf{u}) = \|\mathbf{u}\|_2^2$$

Smooth Graph Signals.



$$x^{(i)} = w^{(i)} + n^{(i)}$$

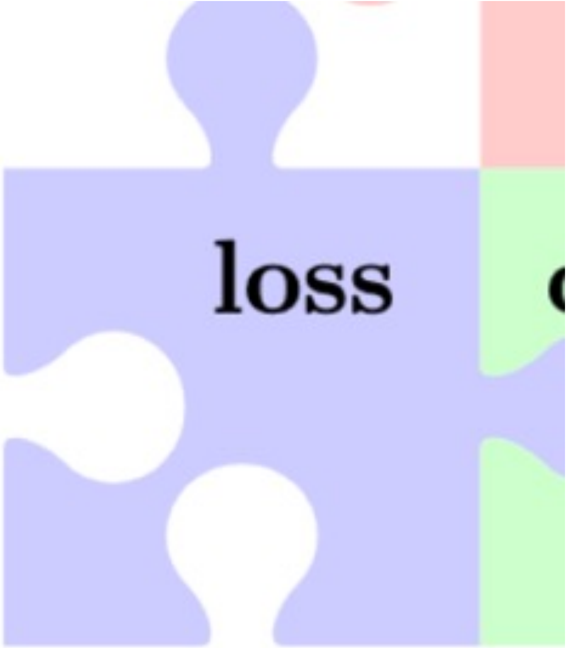
$$A_{i,j}$$

$$x^{(j)} = w^{(j)} + n^{(j)}$$

low pass constraint on

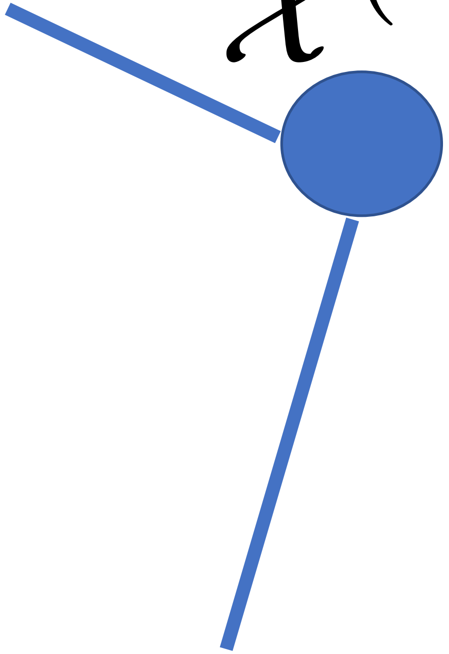
$$\sum_{\{i,j\}} A_{i,j}\left(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\right)^2$$

GTV Minimization.



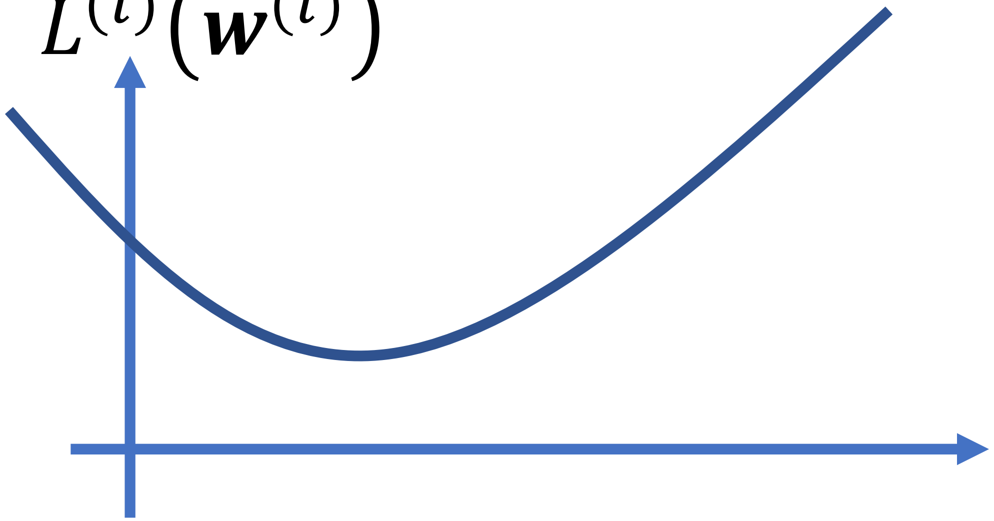
Local Loss Functions.

$\mathcal{X}^{(i)}$



model params $\mathbf{w}^{(i)}$

measure quality of params by local loss function



GTV Minimization.

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$

increasing λ

average local loss

"clusteredness"

training set $\mathcal{M}$

Special Case: Network Lasso.

$$\min_{\mathbf{w}} \sum_{i \in \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|$$

Network Lasso: Clustering and Optimization in Large Graphs

by D Hallac · 2015 · Cited by 206 — **Network Lasso: Clustering and Optimization in Large Graphs** ... Keywords: Convex **Optimization**, ADMM, **Network Lasso.** Go to: ... 2013 [**Google Scholar**]. 2.

Abstract · INTRODUCTION · CONVEX PROBLEM... · EXPERIMENTS

Special Case: “MOCHA”

$$\min_{w} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|^{2}$$

https://papers.nips.cc › paper › 7029-federated-m... ▾ PDF

Federated Multi-Task Learning - NIPS Proceedings

by V Smith · 2017 · Cited by 501 — 3.2 MOCHA: A Framework for **Federated Multi-Task Learning**. In the **federated** setting, the aim is to train statistical models directly on the edge, and thus we solve (1) while assuming that the data {X1,..., Xm} is distributed across m nodes or devices.

Special Case: Graph Sig. Recovery

$$\min_{w} \sum_{i \epsilon \mathrm{M}} \left(x^{(i)} - w^{(i)}\right)^2 + \lambda \sum_{\{i,j\}} A_{i,j}\left(w^{(i)} - w^{(j)}\right)^2$$

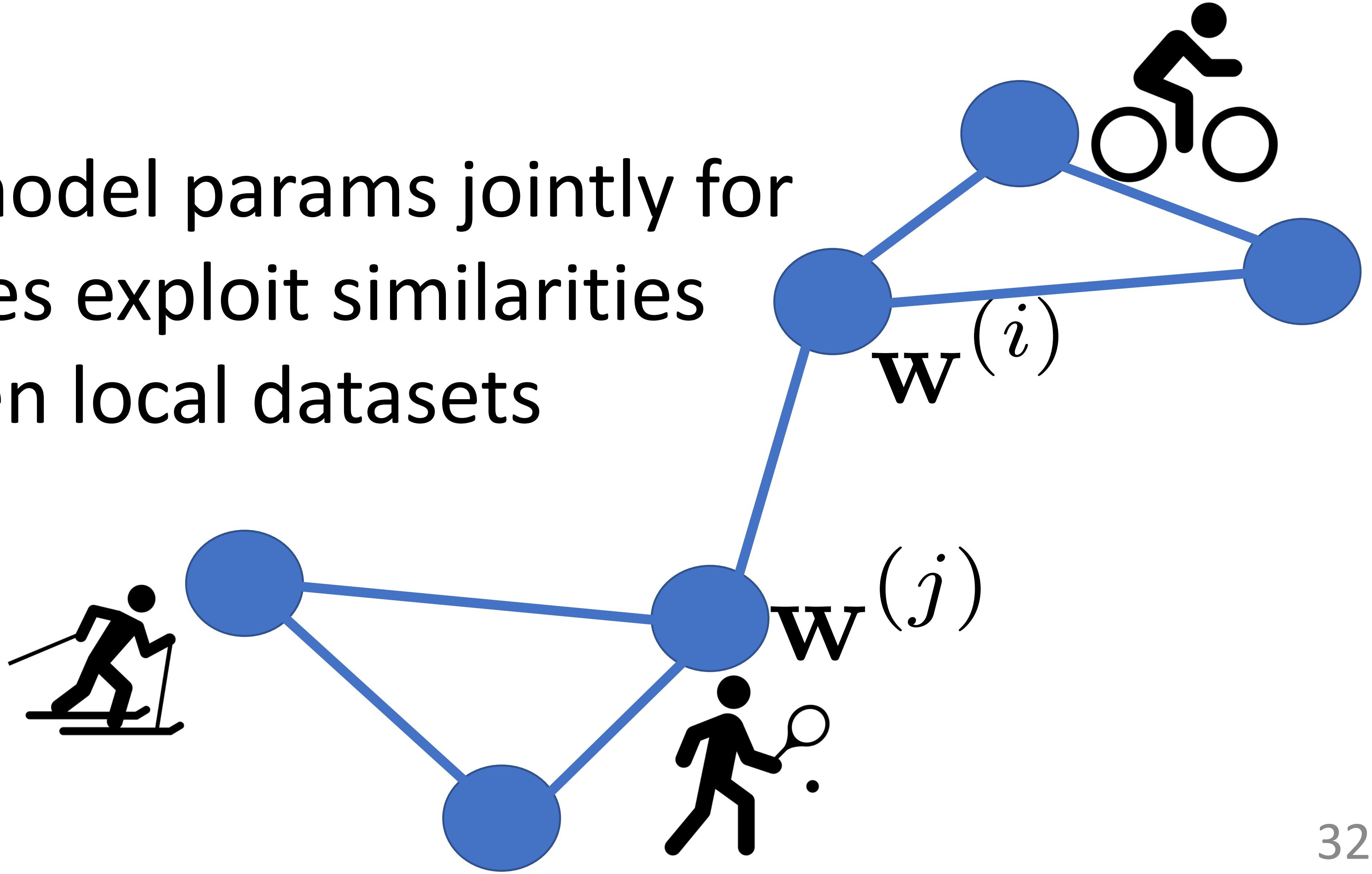
https://papers.nips.cc › paper › 7029-federated-m... ▾ PDF

Federated Multi-Task Learning - NIPS Proceedings

by V Smith · 2017 · Cited by 501 — 3.2 MOCHA: A Framework for **Federated Multi-Task Learning**. In the **federated** setting, the aim is to train statistical models directly on the edge, and thus we solve (1) while assuming that the data {X1,..., Xm} is distributed across m nodes or devices.

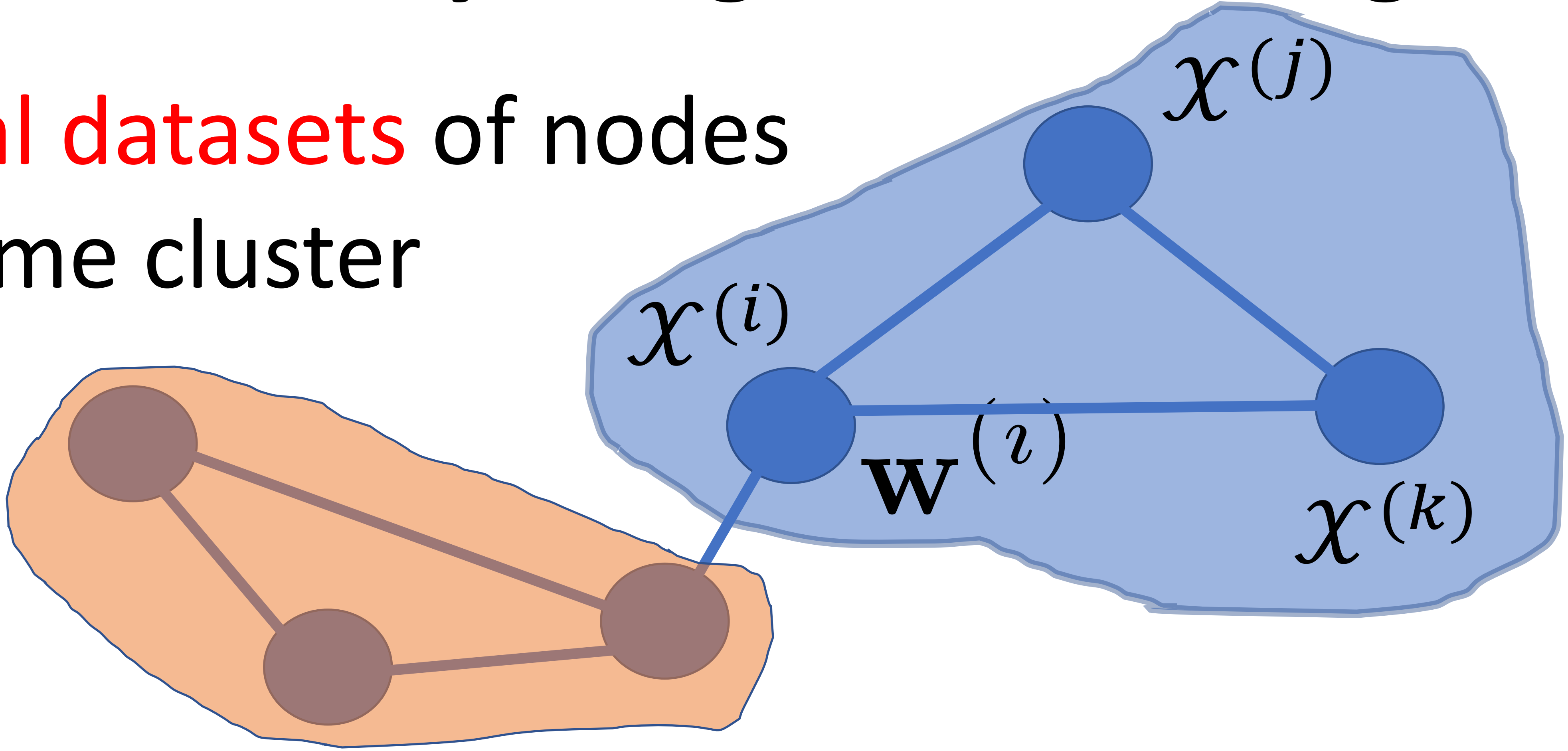
GTVMin is Multi-Task Learning

learn model params jointly for all nodes exploit similarities between local datasets



GTVMin is Locally Weighted Learning

pool local datasets of nodes in the same cluster



William S. Cleveland, Susan J. Devlin, Eric Grosse,
"Regression by local fitting: Methods, properties, and computational algorithms,"
Journal of Econometrics, Volume 37, Issue 1, 1988.

Computational and Statistical Aspects.

how to solve GTVMin efficiently?

are GTVMin solutions statistically useful?

Computational Aspects.

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$

- in-network computation using low-cost devices
- robustness against node/link failures
- robustness against “stragglers”

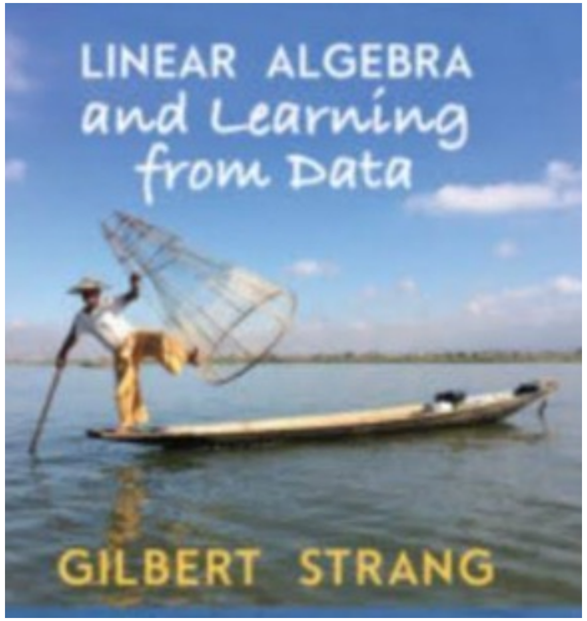
Computational Aspects.

convergence rates; robustness against node failures or “stragglers”; stochastic variants for trading complexity against accuracy

Gradient Descent

$$\underbrace{\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j}\phi\left(w^{(i)} - w^{(j)}\right)}_{f(w)}$$

$$w^{(k+1)} = w^{(k)} - \alpha^{(k)} \nabla f\left(w^{(k)}\right)$$



Iterative Linear Solver.

$$\underbrace{\min_{w} \sum_{i \epsilon \mathrm{M}} \left(x^{(i)} - w^{(i)}\right)^2 + \lambda \sum_{\{i,j\}} A_{i,j}\left(w^{(i)} - w^{(j)}\right)^2}_{f(w)}$$

$$\nabla f(\mathbf{w}) = 0 \;\leftrightarrow\; \mathbf{Lw} = \mathbf{b}$$

Spielman D.A. (2012) Algorithms, Graph Theory, and the Solution of Laplacian Linear Equations. In Automata, Languages, and Programming. ICALP 2012. Lecture Notes in Computer Science, vol 7392. https://doi.org/10.1007/978-3-642-31585-5_5

Rewrite GTVMin using Dual Var.

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$

$\Updownarrow$

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(u^{(i,j)}\right)$$

$$s.t. \mathbf{Bw} = \mathbf{u}$$

Primal-Dual Gradient Method.

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(u^{(i,j)}\right)$$

$$s.t.\ \mathbf{Bw} = \mathbf{u}$$

Algorithm (Incremental PD gradient method)

Setting: Let $J_\rho(w) = J(w) + \frac{\rho}{2}\|Bw - b\|^2$ for some $\rho \geq 0$ and choose positive step-sizes μ_w and μ_λ. Let w_{-1} and λ_{-1} be arbitrary initial conditions and repeat for $i \geq 0$

$$w_i = w_{i-1} - \mu_w\left(\nabla J_\rho(w_{i-1}) + B^\mathsf{T}\lambda_{i-1}\right) \quad (4a)$$

$$\lambda_i = \lambda_{i-1} + \mu_\lambda(Bw_i - b) \quad (4b)$$

S. Alghunaim, A. Sayed, (2020). Linear convergence of primal–dual gradient methods and their performance in distributed optimization. Automatica. 117. 109003. 10.1016/j.automatica.2020.109003.

ADMM (for Non-Smooth Obj.)

$$\min_{\mathbf{w}} \sum_{i \in M} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j}\phi\left(u^{(i,j)}\right)$$

$$s.t.\, \mathbf{Bw} = \mathbf{u}$$

$$x^{k+1} := \underset{x}{\operatorname{argmin}}\, L_\rho(x, z^k, y^k) \qquad (3.2)$$

$$z^{k+1} := \underset{z}{\operatorname{argmin}}\, L_\rho(x^{k+1}, z, y^k) \qquad (3.3)$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c), \qquad (3.4)$$

Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein

Primal Dual Methods

solve GTVMin jointly with its dual!

dual of GTVMin has remarkable interpretation…

Primal Form of GTVMin.

$$\min_{\mathbf{w}} f(\mathbf{w}) + g\big(\mathbf{D}\mathbf{w}\big)$$

$$f(\mathbf{w}) := \sum_{\boldsymbol{i} \,\boldsymbol{\epsilon}\, \mathbf{M}} L^{(\boldsymbol{i})}\big(\boldsymbol{w}^{(\boldsymbol{i})}\big) \qquad g(\mathbf{u}) := \lambda \sum_{e \in \mathcal{E}} A_e \phi\big(\mathbf{u}^{(e)}\big)$$

primal variables $\mathbf{w} : \mathcal{V} \rightarrow \mathbb{R}^n : i \mapsto \mathbf{w}^{(i)}$

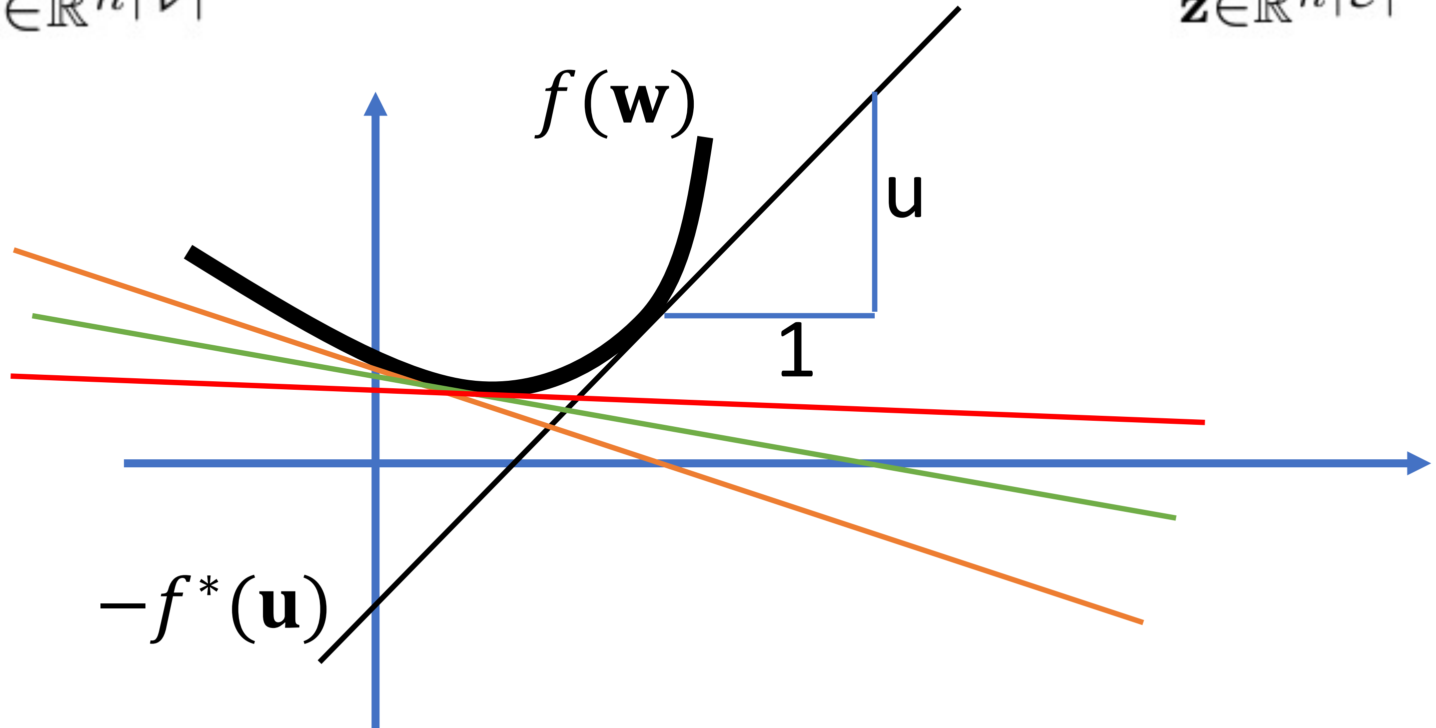
dual variables $\mathbf{u} : \mathcal{E} \rightarrow \mathbb{R}^n : e \mapsto \mathbf{u}^{(e)}$

block-incidence matrix $\mathbf{D} \in \{-1, 1, 0\}^{\mathcal{E} \times \mathcal{V}}$

Dual of GTVMin.

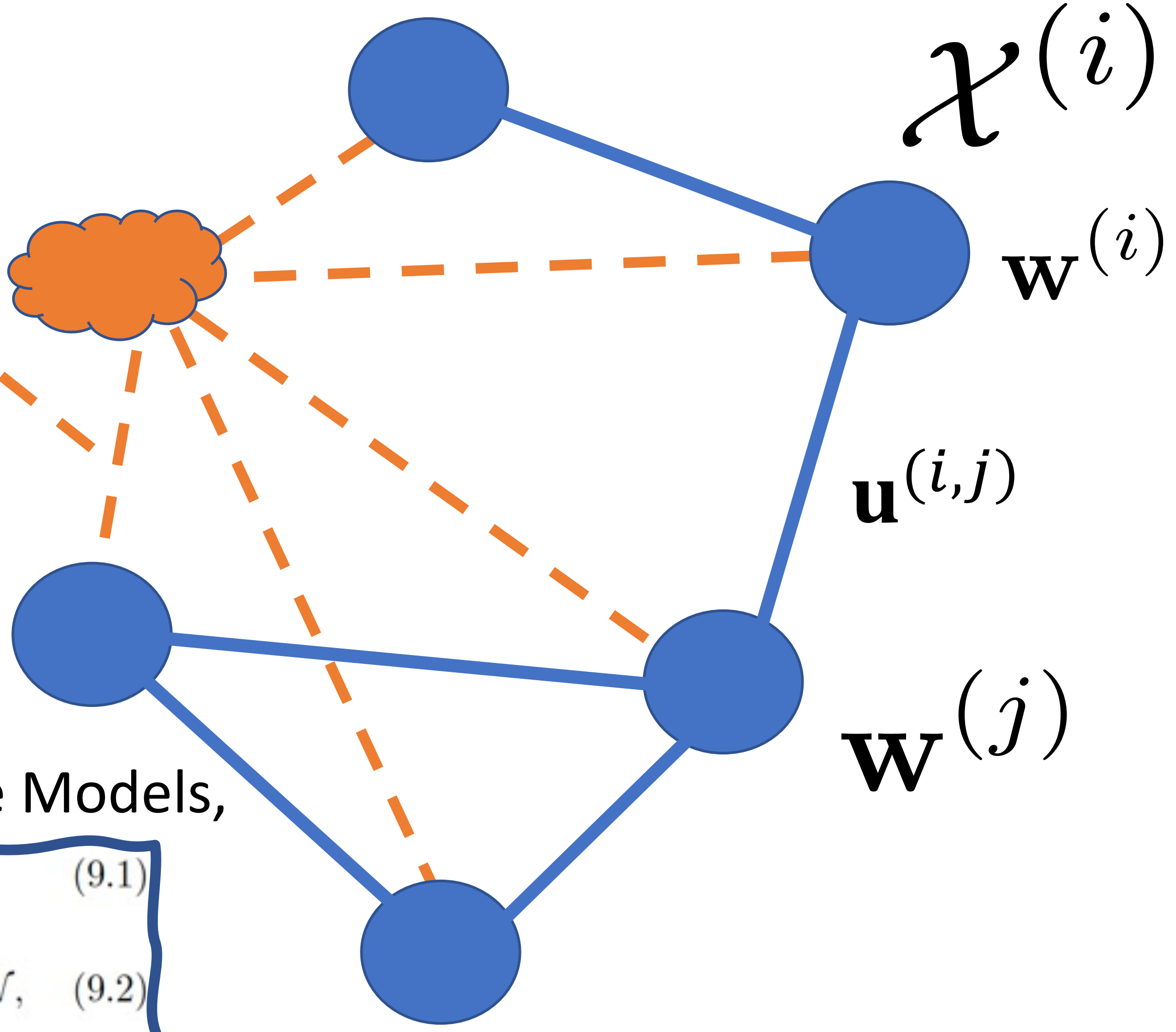
$$\max_{\mathbf{u}\in\mathbb{R}^{n|\mathcal{E}|}} -g^*(\mathbf{u}) - f^*(-\mathbf{D}^T\mathbf{u}).$$

$$f^*(\mathbf{w}) := \sup_{\mathbf{z}\in\mathbb{R}^{n|\mathcal{V}|}} \mathbf{w}^T\mathbf{z} - f(\mathbf{z}) \qquad g^*(\mathbf{u}) := \sup_{\mathbf{z}\in\mathbb{R}^{n|\mathcal{E}|}} \mathbf{u}^T\mathbf{z} - g(\mathbf{z})$$



Dual of GTVMin = Min. Cost Flow

$$\max_{\mathbf{u}\in\mathbb{R}^{n|\mathcal{E}|}} -g^*(\mathbf{u}) - f^*(-\mathbf{D}^T\mathbf{u}).$$



D. Bertsekas,
Network Optimization: Continuous and Discrete Models,
1998.

$$
\begin{aligned}
&\text{minimize} && \sum_{(i,j)\in\mathcal{A}} f_{ij}(x_{ij}) && (9.1)\\
&\text{subject to} && \sum_{\{j|(i,j)\in\mathcal{A}\}} x_{ij} - \sum_{\{j|(j,i)\in\mathcal{A}\}} x_{ji} = s_i, \quad \forall\, i \in \mathcal{N}, && (9.2)\\
& && x_{ij} \in X_{ij}, \quad \forall\, (i,j) \in \mathcal{A}, && (9.3)
\end{aligned}
$$

Primal-Dual Optimality Conditions.

(assuming convexity of loss functions and GTV)

primal and dual variables $\widehat{\mathrm{w}}, \hat{\mathrm{u}}$ optimal if and only if

$$\mathbf{M}^{-1}\begin{pmatrix} \partial f & \mathbf{D}^{T} \\ -\mathbf{D} & \partial g^{*} \end{pmatrix}\begin{pmatrix} \widehat{\mathbf{w}} \\ \widehat{\mathbf{u}} \end{pmatrix} \ni \mathbf{0} \text{ with } \mathbf{M} := \begin{pmatrix} \mathbf{T}^{-1} & -\mathbf{D}^{T} \\ -\mathbf{D} & \mathbf{\Sigma}^{-1} \end{pmatrix}$$

$$\left(\mathbf{\Sigma}\right)_{e,e} := \sigma_e \mathbf{I}_n, \text{ for } e \in \mathcal{E}, \left(\mathbf{T}\right)_{i,i} := \tau_i \mathbf{I} \text{ for } i \in \mathcal{V},$$

$$\text{with } \sigma_e := 1/2 \text{ for } e \in \mathcal{E} \text{ and } \tau_i := 1/|\mathcal{N}_i| \text{ for } i \in \mathcal{V}.$$

R. T. Rockafellar , CONVEX ANALYSIS, Princeton Univ. Press, 1970.

Proximal Point Algorithm.

primal and dual variables $\widehat{\mathrm{w}}, \hat{\mathrm{u}}$ optimal if and only if

$$\mathbf{M}^{-1}\begin{pmatrix}\partial f & \mathbf{D}^T \\ -\mathbf{D} & \partial g^*\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{w}} \\ \widehat{\mathbf{u}}\end{pmatrix} \ni \mathbf{0} \text{ with } \mathbf{M} := \begin{pmatrix}\mathbf{T}^{-1} & -\mathbf{D}^T \\ -\mathbf{D} & \mathbf{\Sigma}^{-1}\end{pmatrix}$$

solve iteratively by proximal point algorithm

$$\begin{pmatrix}\widehat{\mathbf{w}}^{(k+1)} \\ \widehat{\mathbf{u}}^{(k+1)}\end{pmatrix} = \left(\mathbf{I} + \mathbf{M}^{-1}\begin{pmatrix}\partial f & \mathbf{D}^T \\ -\mathbf{D} & \partial g^*\end{pmatrix}\right)^{-1}\begin{pmatrix}\widehat{\mathbf{w}}^{(k)} \\ \widehat{\mathbf{u}}^{(k)}\end{pmatrix}$$

A. Chambolle, T. Pock. An introduction to continuous optimization for imaging. Acta Numerica, 2016.

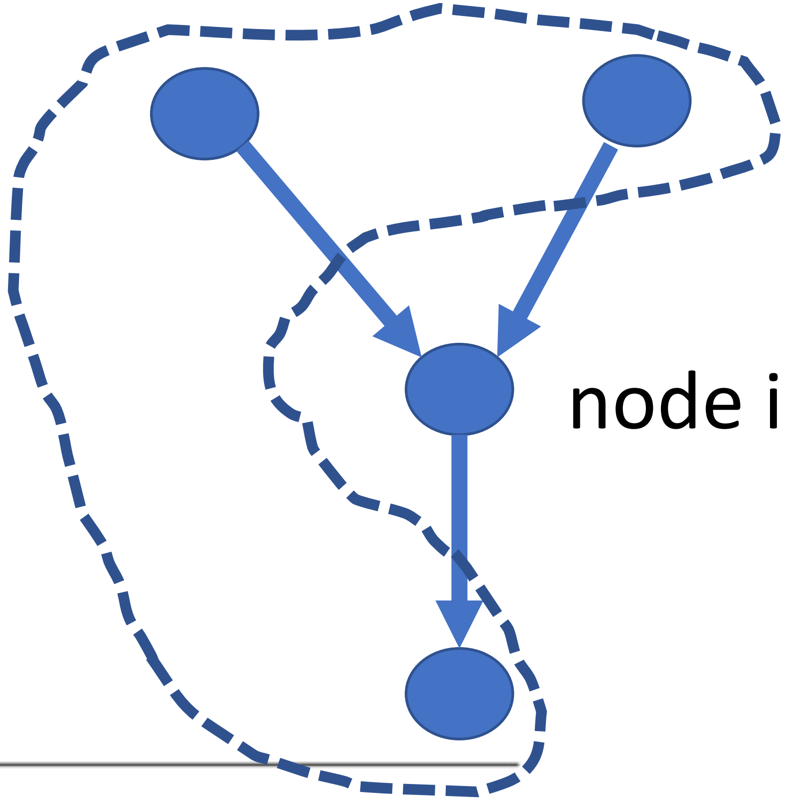
After Some Manipulations.

Algorithm 1 Primal-Dual Method for Networked FL

Input: empirical graph $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathbf{A})$; training set $\{\mathbf{X}^{(i)}\}_{i\in\mathcal{M}}$; regularization parameter λ; loss $\mathcal{L}$; GTV penalty ϕ

Initialize: $k:=0; \widehat{\mathbf{w}}_0:=\mathbf{0}; \widehat{\mathbf{u}}_0:=\mathbf{0}$; $\sigma_e = 1/2$ and $\tau_i = 1/|\mathcal{N}_i|$

while stopping criterion is not satisfied **do**
    for all nodes $i \in \mathcal{V}$ **do**
        $\widehat{\mathbf{w}}^{(i)}_{k+1} := \widehat{\mathbf{w}}^{(i)}_{k} - \tau_i \sum_{e\in\mathcal{E}} D_{e,i} \widehat{\mathbf{u}}^{(e)}_{k}$
    end for
    for nodes in the training set $i \in \mathcal{M}$ **do**
        $\widehat{\mathbf{w}}^{(i)}_{k+1} := \mathcal{PU}^{(i)}\{\widehat{\mathbf{w}}^{(i)}_{k+1}\}$
    end for
    for all edges $e \in \mathcal{E}$ **do**
        $\widehat{\mathbf{u}}^{(e)}_{k+1} := \widehat{\mathbf{u}}^{(e)}_{k} + \sigma_e\big(2\big(\widehat{\mathbf{w}}^{(e_+)}_{k+1} - \hat{\mathbf{w}}^{(e_-)}_{k+1}\big) - \big(\widehat{\mathbf{w}}^{(e_+)}_{k} - \widehat{\mathbf{w}}^{(e_-)}_{k}\big)\big)$
        $\widehat{\mathbf{u}}^{(e)}_{k+1} := \mathcal{DU}^{(e)}\{\widehat{\mathbf{u}}^{(e)}_{k+1}\}$
    end for
    $k:=k+1$
end while

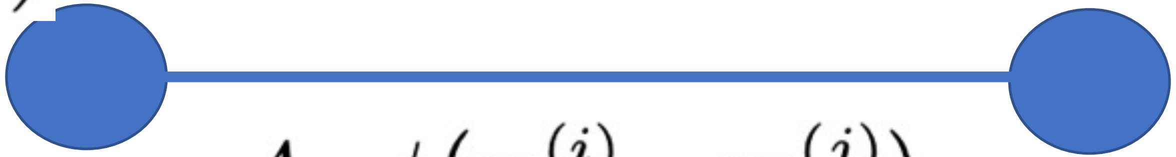


Algorithm 1 is Attractive for NFL…

- decentralized implementation (mess. pass.)
- robust against various imperfections
  - approximate primal/dual updates
  - node/link failures
- privacy friendly; no raw data exchanged

Local Computations in Algorithm 1.

$$L^{(i)}\left(\mathcal{X}^{(i)}, \mathbf{w}^{(i)}\right)$$

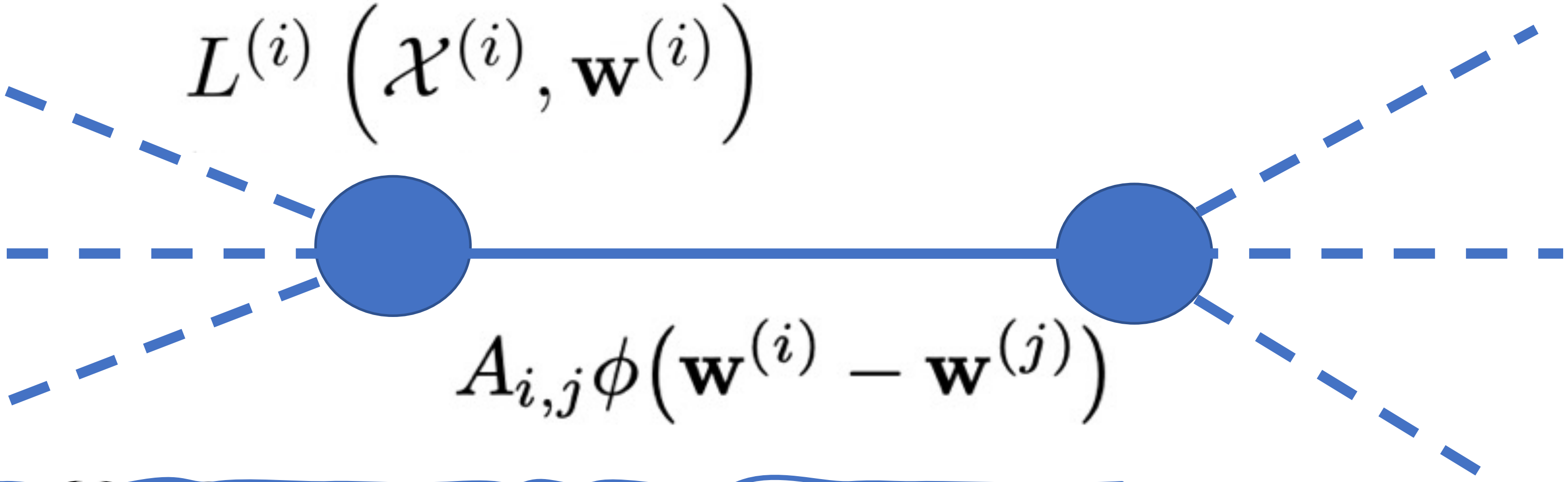


$$A_{i,j}\phi\big(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\big)$$

node-wise
primal update:
$$\mathcal{PU}^{(i)}\{\mathbf{v}\} := \underset{\mathbf{z}\in\mathbb{R}^n}{\operatorname{argmin}}\, L^{(i)}\left(\mathbf{z}\right)+(1/2\tau_i)\|\mathbf{v}-\mathbf{z}\|^2.$$

edge-wise
dual update:
$$\mathcal{DU}^{(e)}\{\mathbf{v}\} := \underset{\mathbf{z}\in\mathbb{R}^n}{\operatorname{argmin}}\, \lambda A_e\phi^*\left(\mathbf{z}/(\lambda A_e)\right)+(1/2\sigma_e)\|\mathbf{v}-\mathbf{z}\|^2.$$

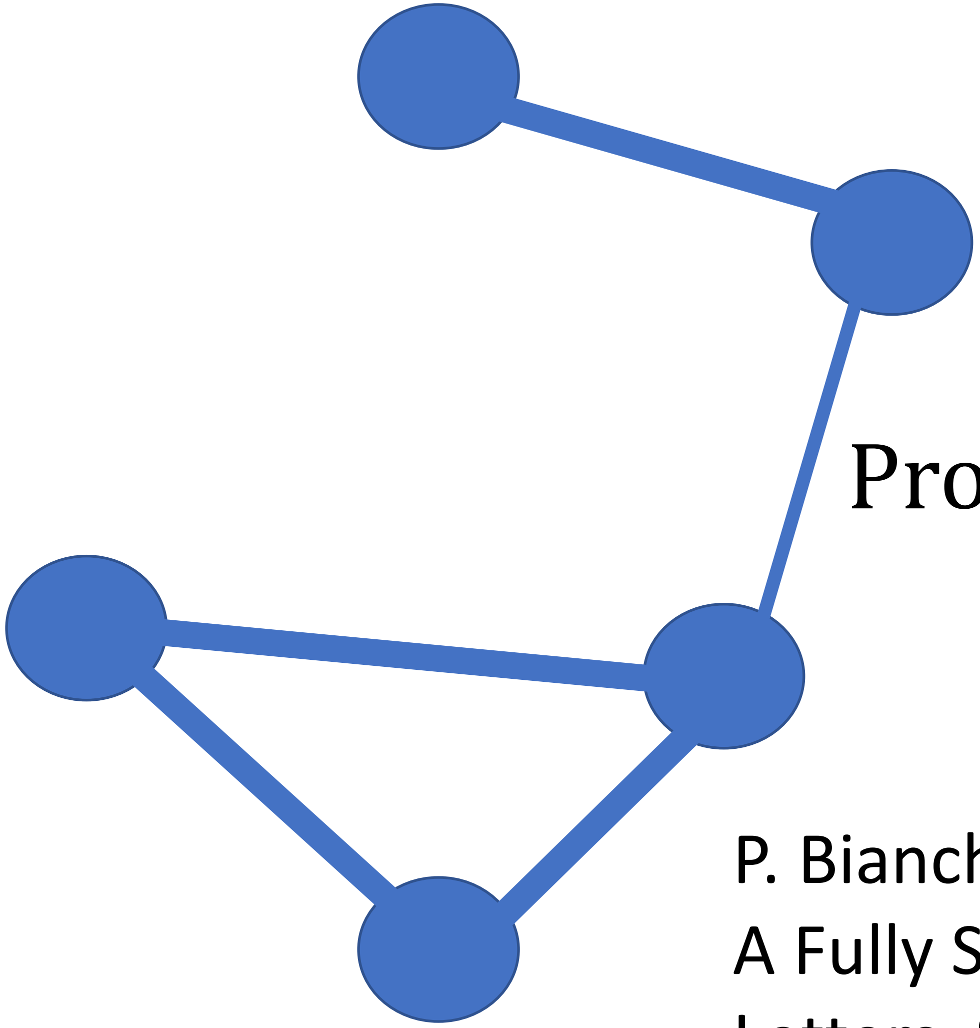
Spreading Local Results.



for all nodes $i \in \mathcal{V}$ **do**
    $\widehat{\mathbf{w}}^{(i)}_{k+1} := \widehat{\mathbf{w}}^{(i)}_{k} - \tau_i \sum_{e \in \mathcal{E}} D_{e,i} \widehat{\mathbf{u}}^{(e)}_{k}$
end for

for all edges $e \in \mathcal{E}$ **do**
    $\widehat{\mathbf{u}}^{(e)}_{k+1} := \widehat{\mathbf{u}}^{(e)}_{k} + \sigma_e\big(2\big(\widehat{\mathbf{w}}^{(e_+)}_{k+1} - \widehat{\mathbf{w}}^{(e_-)}_{k+1}\big) - \big(\widehat{\mathbf{w}}^{(e_+)}_{k} - \widehat{\mathbf{w}}^{(e_-)}_{k}\big)\big)$

Networked Data as Realizations of RV

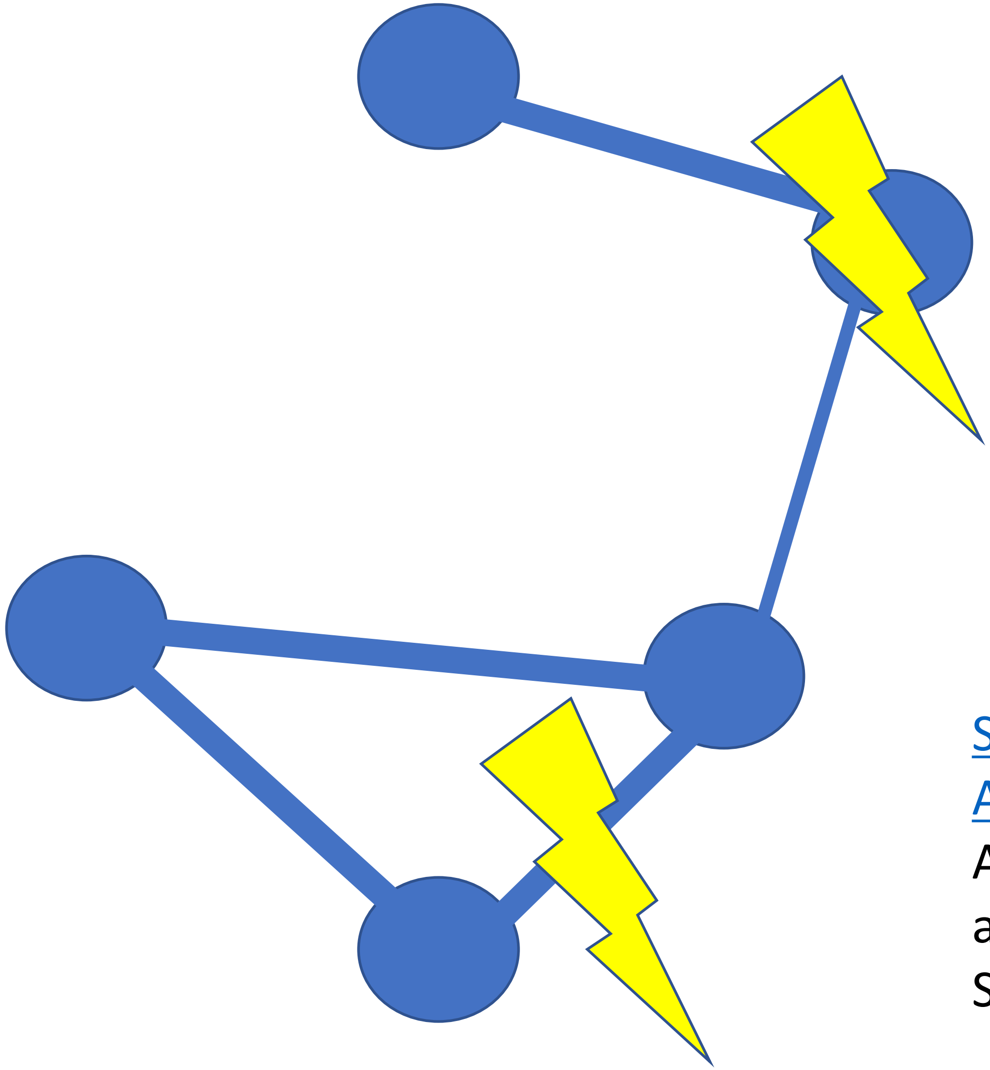


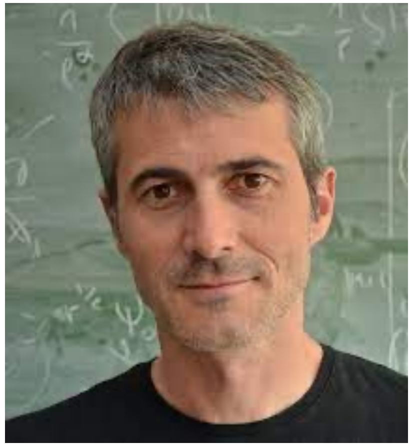
$$\mathcal{X}^{(i)} \ \text{iid} \ \sim p^{(i)}(\mathbf{z})$$

$$\text{Prob}(A_{i,j} = 1) = p$$

P. Bianchi, W. Hachem, A. Salim.
A Fully Stochastic Primal-Dual Algorithm. Optimization Letters, Springer Verlag, 2020,

Random Node/Link Failures.



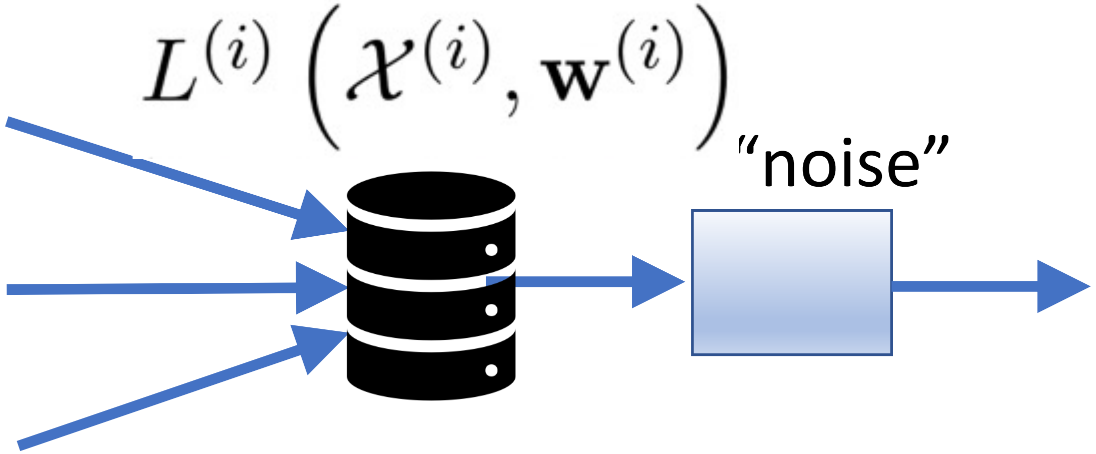


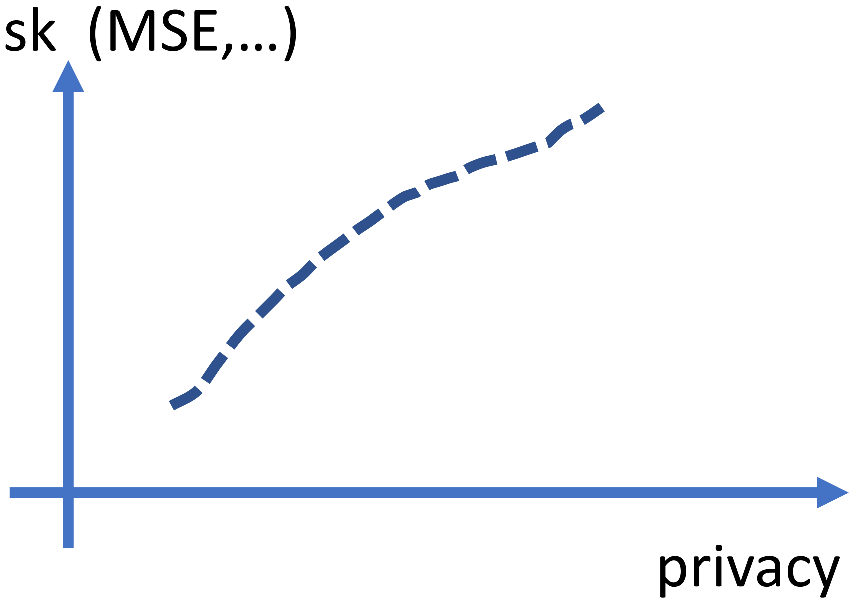
[Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications]

Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb

SIAM Journal on Optimization 2018 28:4, 2783-2808

Privacy-Preservation.





- Huang, Z. and Gong, Y., “Differentially Private ADMM for Convex Distributed Learning: Improved Accuracy via Multi-Step Approximation”, <i>arXiv e-prints</i>, 2020.
- Huang, Z., Hu, R., Guo, Y., Chan-Tin, E., and Gong, Y., “DP-ADMM: ADMM-based Distributed Learning with Differential Privacy”, <i>arXiv e-prints</i>, 2018.
- J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in Proc. IEEE Annu. Symp. Found. Comput. Sci., pp. 429–438, 2013.

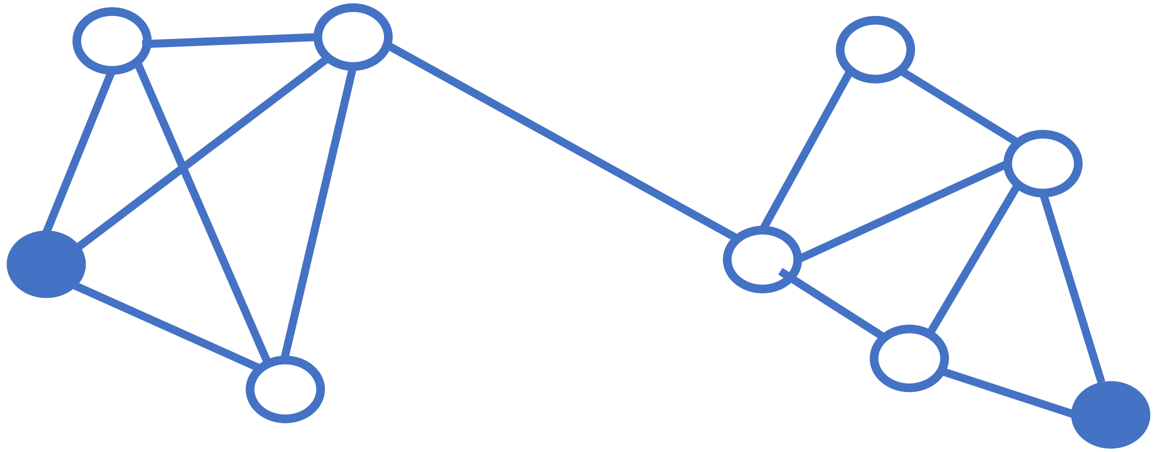
Bottom Line.

established distributed optimization provides efficient technology for solving GTVMin in robust and privacy-friendly way

....., however ....

Are GTVMin Solutions Any Good?

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$



training/sampling set $\mathcal{M}$

which combination of signal model (choice of ϕ) and sampling set M ensure solutions of GTVMin are "sensible" ?

Statistical Aspects.

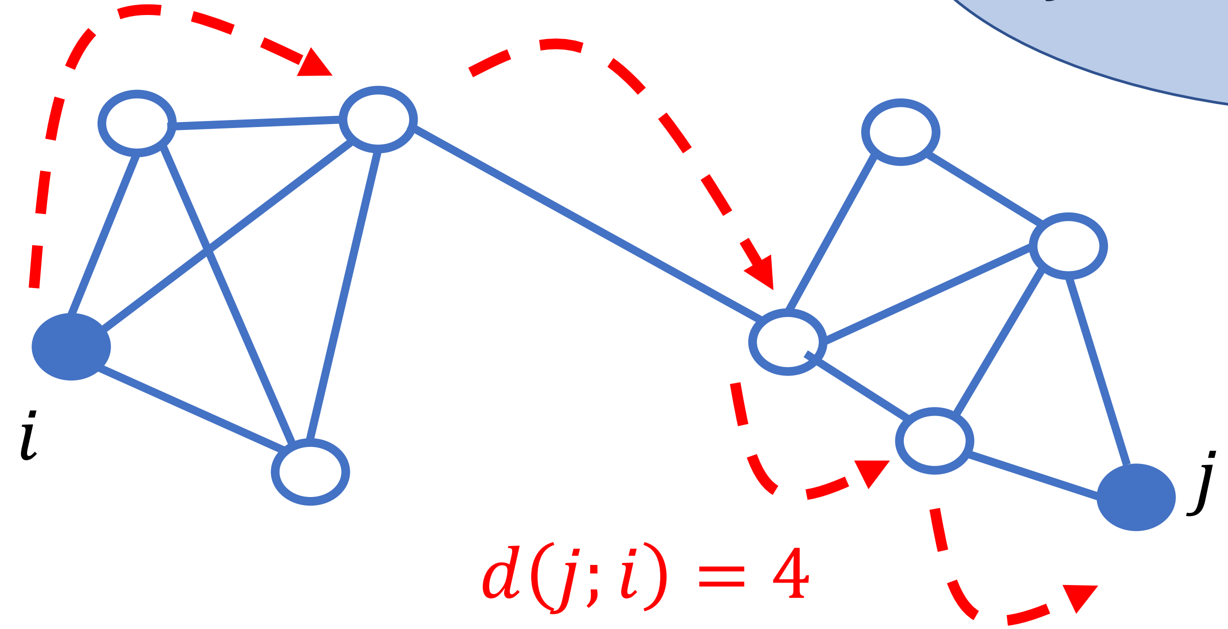
$$min_w \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$

statistical properties of GTVMin solutions?

- sampling theorems (signal processing)
- generalization bounds (ML perspective)

Signal Processing Perspective.

$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} \left(w^{(i)} - y^{(i)}\right)^2 + \lambda \sum_{\{i,j\}} A_{i,j}\left(w^{(i)} - w^{(j)}\right)^2$$



constraints
effective "bandwidth"
of graph signal $w^{(i)}$

M. Tsitsvero, S. Barbarossa and P. Di Lorenzo, "Signals on Graphs: Uncertainty Principle and Sampling," in *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4845-4860, 15 Sept.15, 2016, doi: 10.1109/TSP.2016.2573748.

Machine Learning Perspective.

Theorem 1 (Generalization Performance of Graph Regularization). *Let γ be the regularization parameter, T be a set of $k \geq 4$ vertices $\mathbf{x}_1, \ldots, \mathbf{x}_k$, where each vertex occurs no more than t times, together with values $y_1, \ldots, y_k$, $|y_i| \leq M$. Let f_T be the regularization solution using the smoothness functional S with the second smallest eigenvalue λ_1. Assuming that $\forall \mathbf{x} |f_T(\mathbf{x})| \leq K$ we have with probability $1 - \delta$ (conditional on the multiplicity being no greater than t):*

$$|R_k(f_T) - R(f_T)| \leq \beta + \sqrt{\frac{2\log(2/\delta)}{k}}\,(k\beta + (K+M)^2)$$

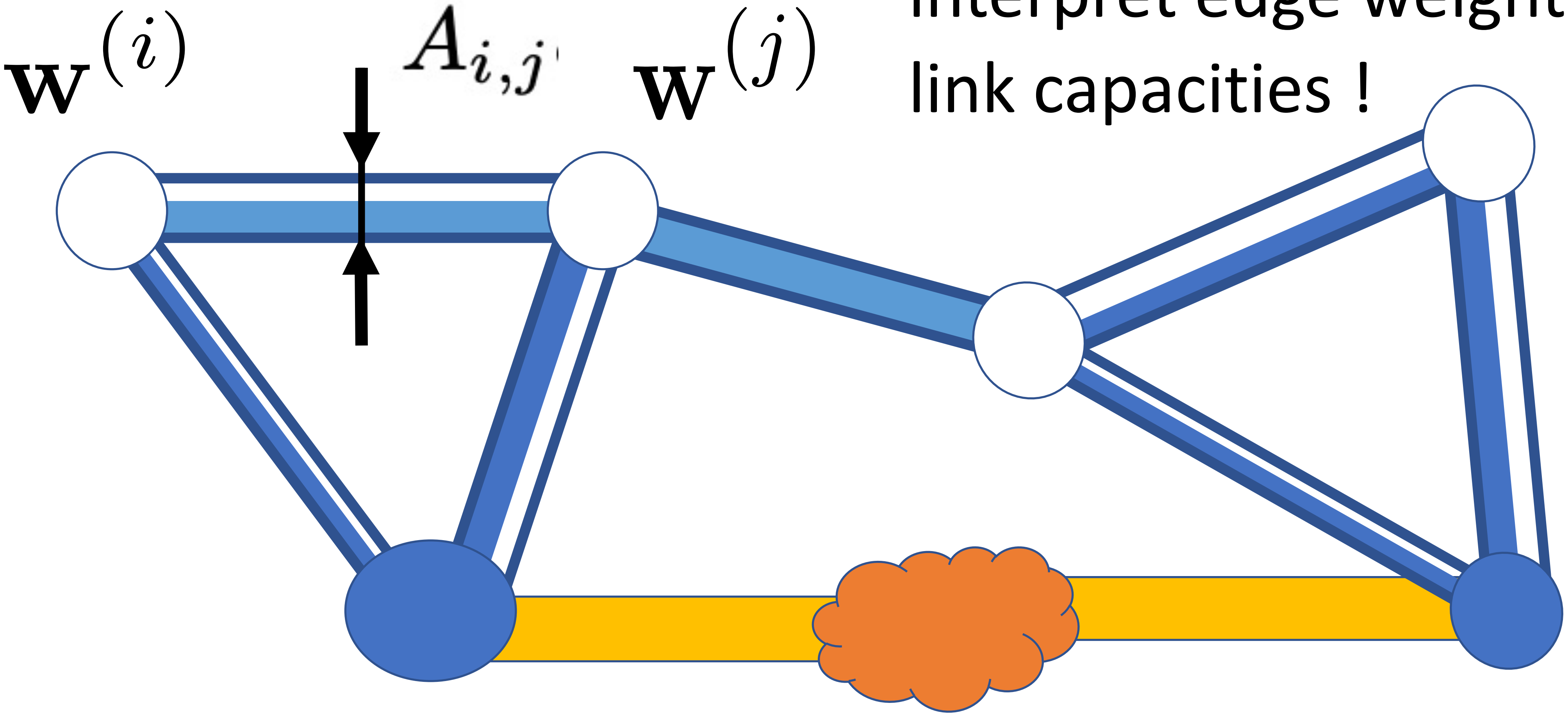
where

$$\beta = \frac{3M\sqrt{tk}}{(k\gamma\lambda_1 - t)^2} + \frac{4M}{k\gamma\lambda_1 - t}$$

Belkin M., Matveeva I., Niyogi P. Regularization and Semi-supervised Learning on Large Graphs. COLT 2004. Lecture Notes in Computer Science, Springer, 2004
https://doi.org/10.1007/978-3-540-27819-1_43

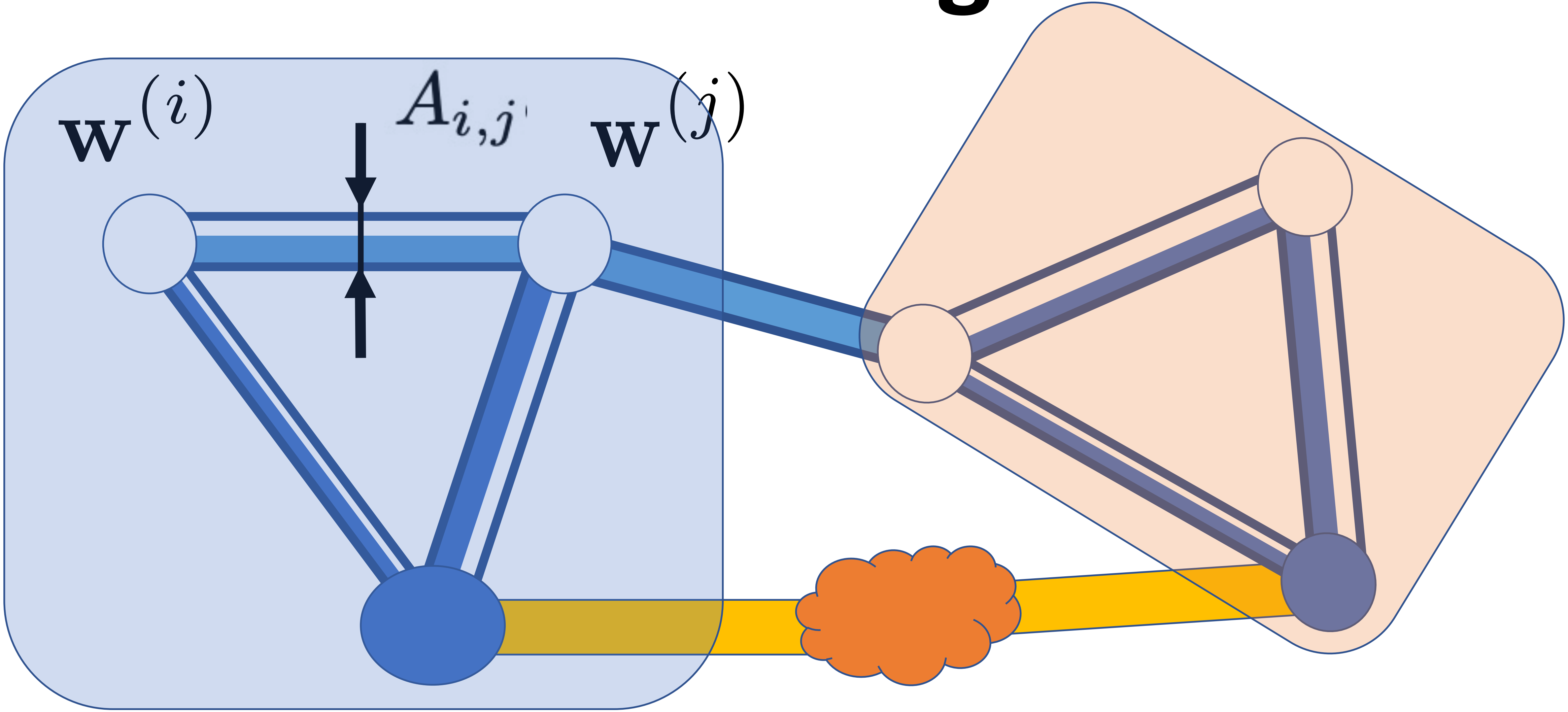
Our Perspective: Flows.

interpret edge weights $A_{i,j}$ as link capacities !



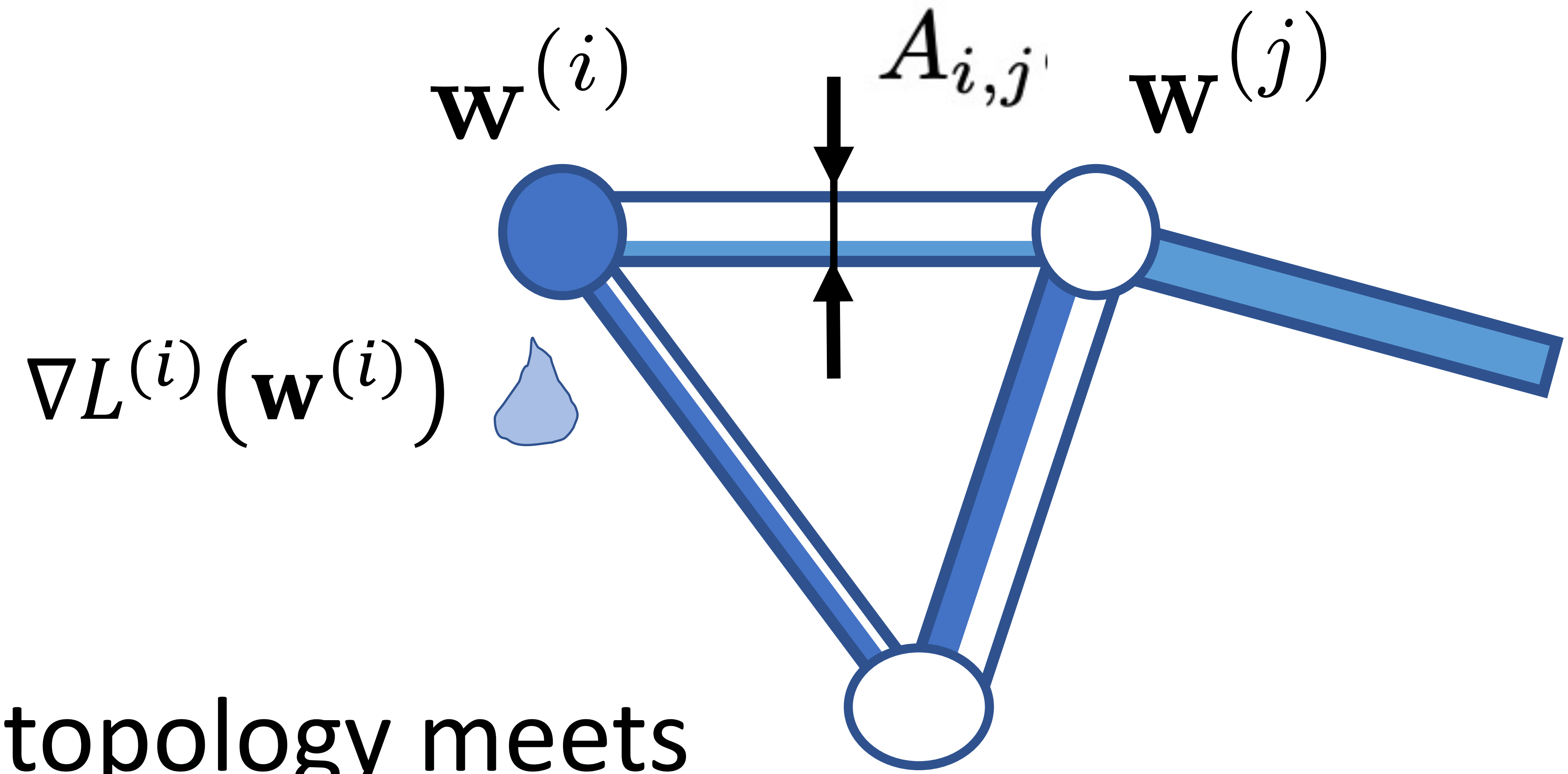
A. Jung, "On the Duality Between Network Flows and Network Lasso," in *IEEE Signal Processing Letters*, vol. 27, pp. 940-944, 2020.

Cluster-wise Pooling.



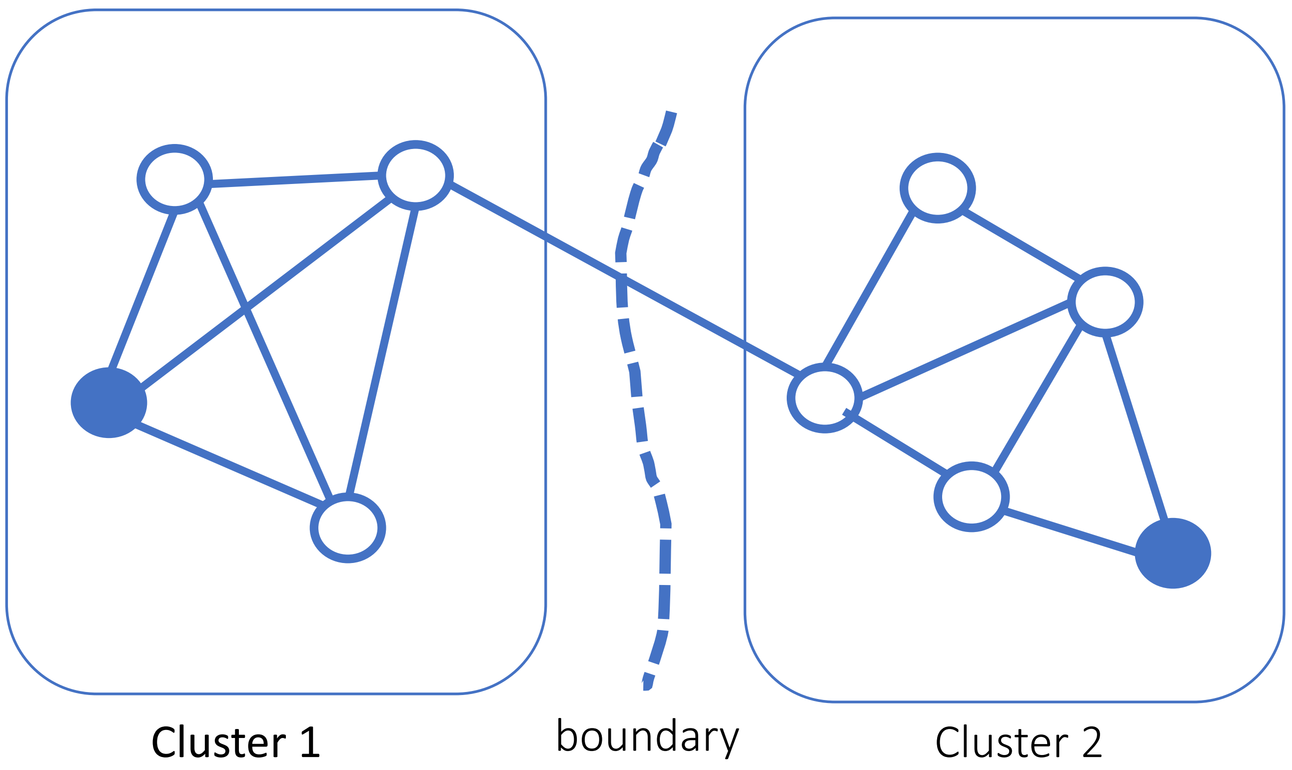
parameter vectors can only change over statured links

Leaky Training Set.



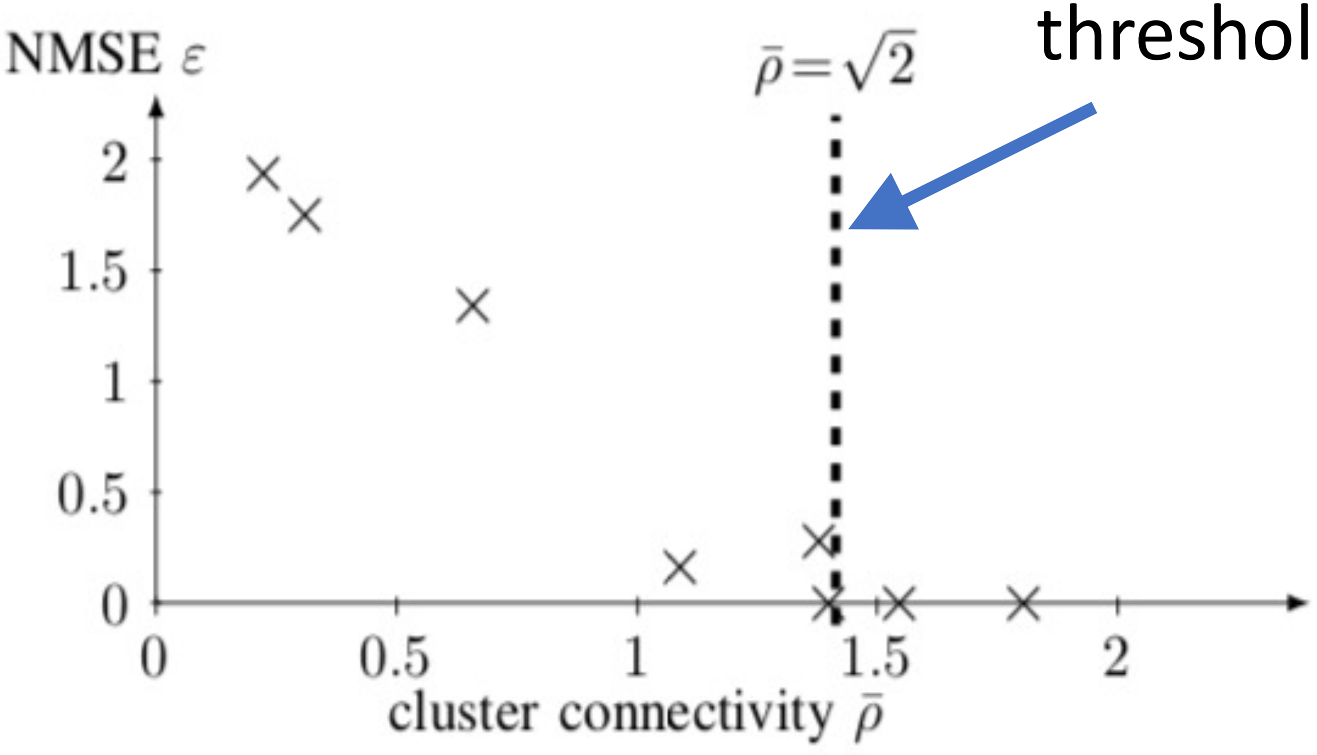
network topology meets
geometry of loss functions !

Measure Connectivity by Flows.



connectivity measured by flow ρ that can be routed over boundary edge

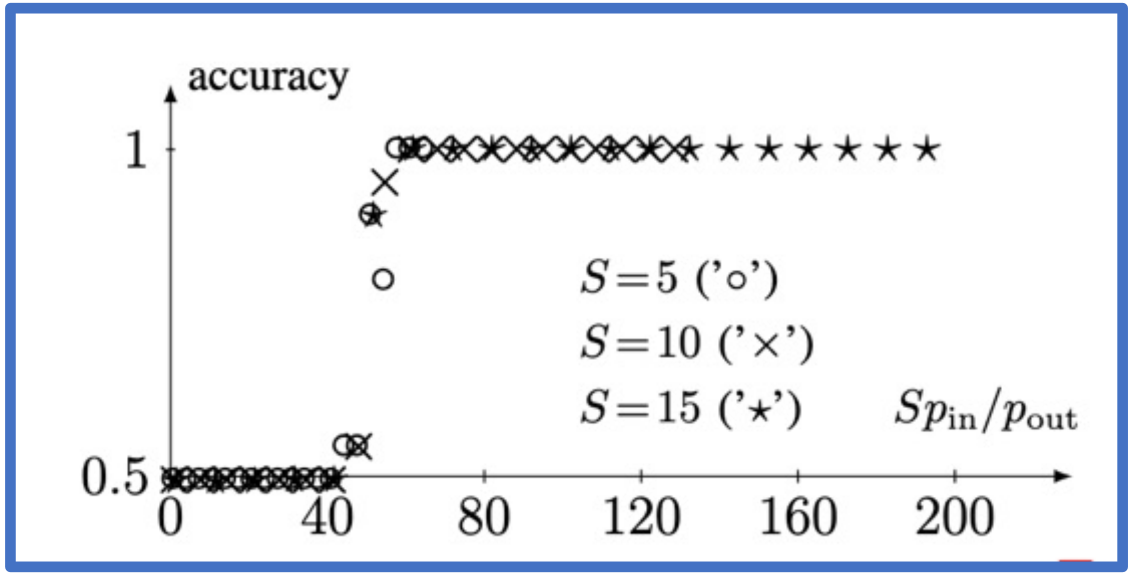
Statistical Error vs. Connectivity.



threshold predicted by our analysis

A. Jung and N. Tran, "Localized Linear Regression in Networked Data," in *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1090-1094, July 2019.

Clustering Assumption in SBM.

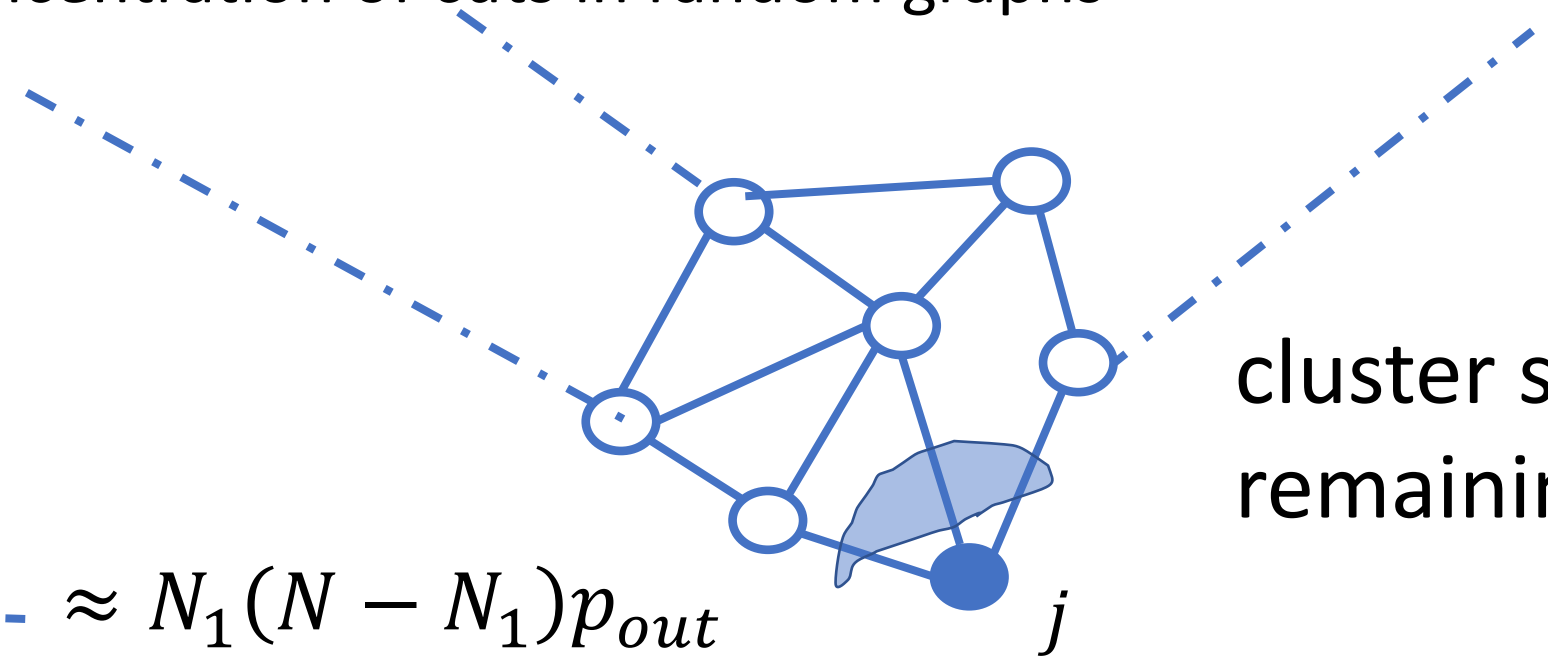


- intra-cluster edge prob p_{in}
- inter-cluster edge prob p_{out}
- S training nodes in each cluster
- critical value for S*pin/pout

A. Jung,
"Clustering in Partially Labeled Stochastic Block Models via Total Variation Minimization,"
54th Asilomar Conference on Signals, Systems, and Computers, 2020,

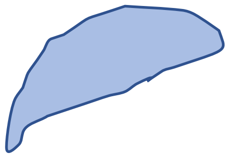
Mathematical Device.

- flow conservation/Hoffman's circulation theorem
- concentration of cuts in random graphs



cluster size N_1
remaining nodes $N - N_1$

$\approx N_1(N - N_1)p_{out}$



$\approx N_1 p_{in}$

R. Karger,
Random sampling in cut, flow, and network design problems,
Math. Oper. Res., 24 (1999), pp. 383–413.

Wrap Up.

- formulated federated learning as GTV minimization
- two special cases: network Lasso and MOCHA
- solved GTV min. with established primal-dual method
- scalable and robust implementation as message passing
- GTV min. adaptively pools similar datasets

Want to dig deeper ?

upcoming

IEEE SPS Seasonal School on Networked Federated Learning

Tentative Schedule

Each day consists of lectures and exercises.

Mo. 28.03.	Tue. 29.03.	Wed. 30.03.	Th. 31.03.	Fr. 01.04.
Machine Learning	**Networks**	**Basic FL**	**Clustered FL**	**Trustworthy FL**
Data, Model, Loss	Graphs and their Matrices	Networked Data	Networked Models	Privacy-Preservation
Linear and Logistic Regression	Spectrum of Laplacian	Centralized FL	Total Variation Minimization	Explainability
Gradient-Based Learning	Cluster Structure	Gossip, Consensus	Distributed SGD	Legal Aspects

https://ieeespcasfinland.github.io/

Thank you for
your attention!