Networked Federated Learning

Alexander Jung (Aalto University)

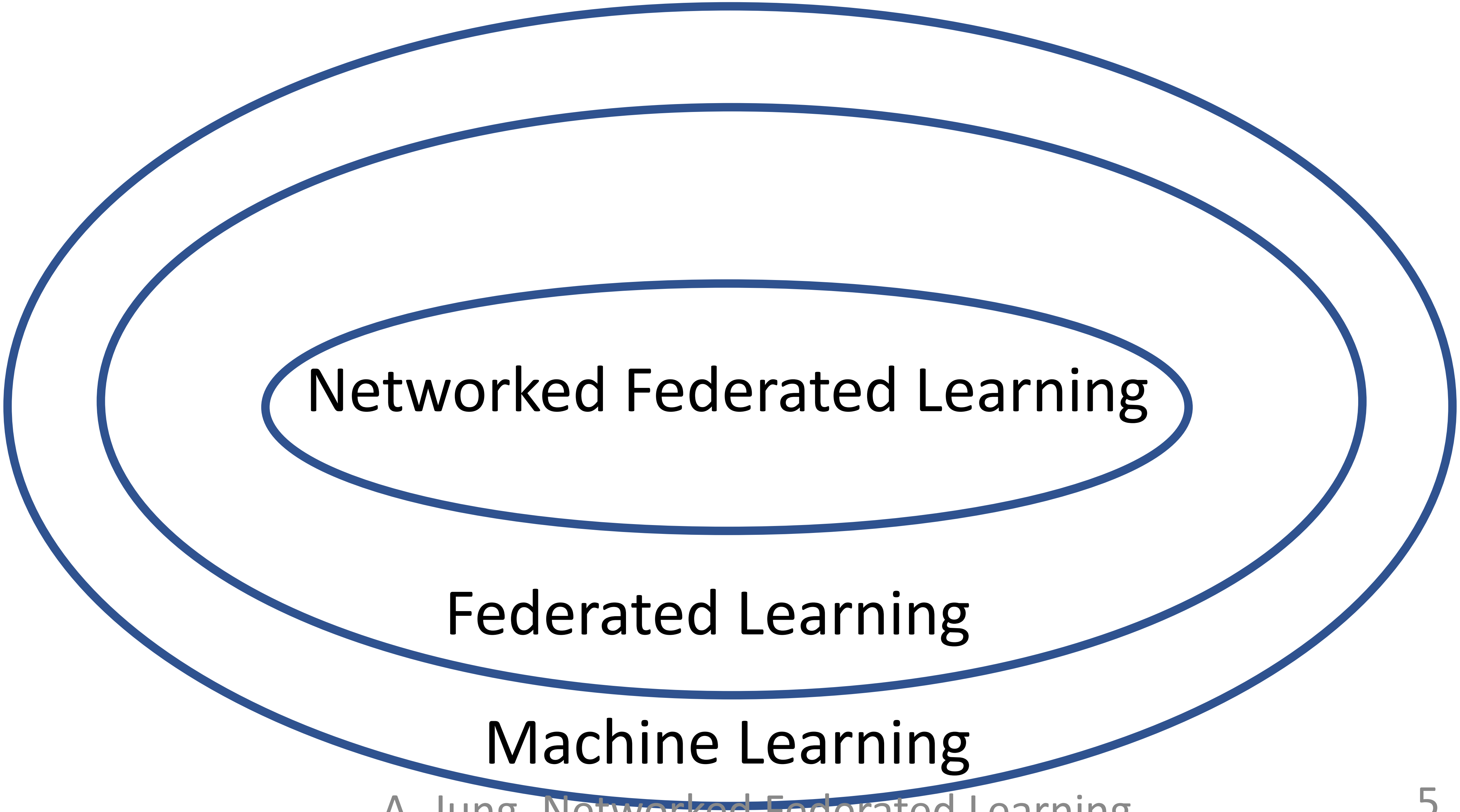
https://www.linkedin.com/in/aljung/
https://www.youtube.com/channel/UC_tW4Z_GfJ2WCnKDtwMuDUA
https://twitter.com/alexjungaalto

- GTVMin as NFL Principle
- The Dual of GTVMin
- Interpretations
- Computational Aspects
- Statistical Aspects

- GTVMin as NFL Principle
- The Dual of GTVMin
- Interpretations
- Computational Aspects
- Statistical Aspects

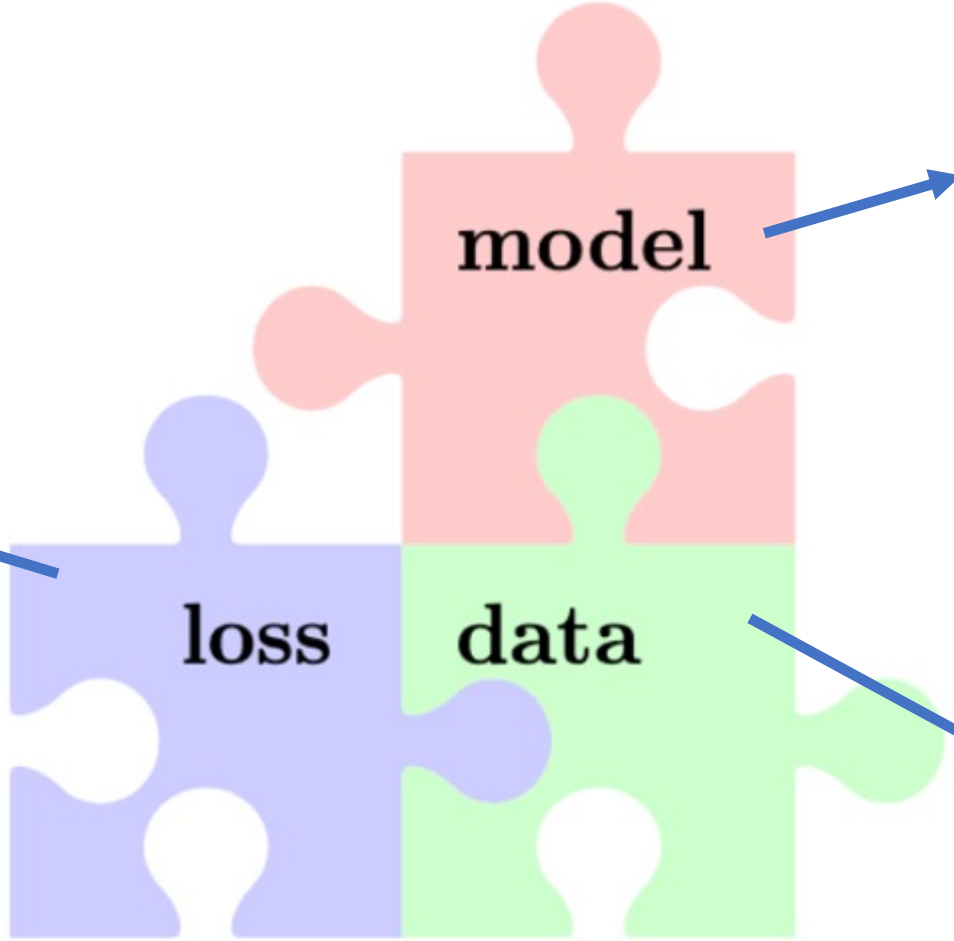
In a nutshell:

organize data, models and computation for machine learning as networks.


Networked Federated Learning
Federated Learning
Machine Learning

Three Components of ML

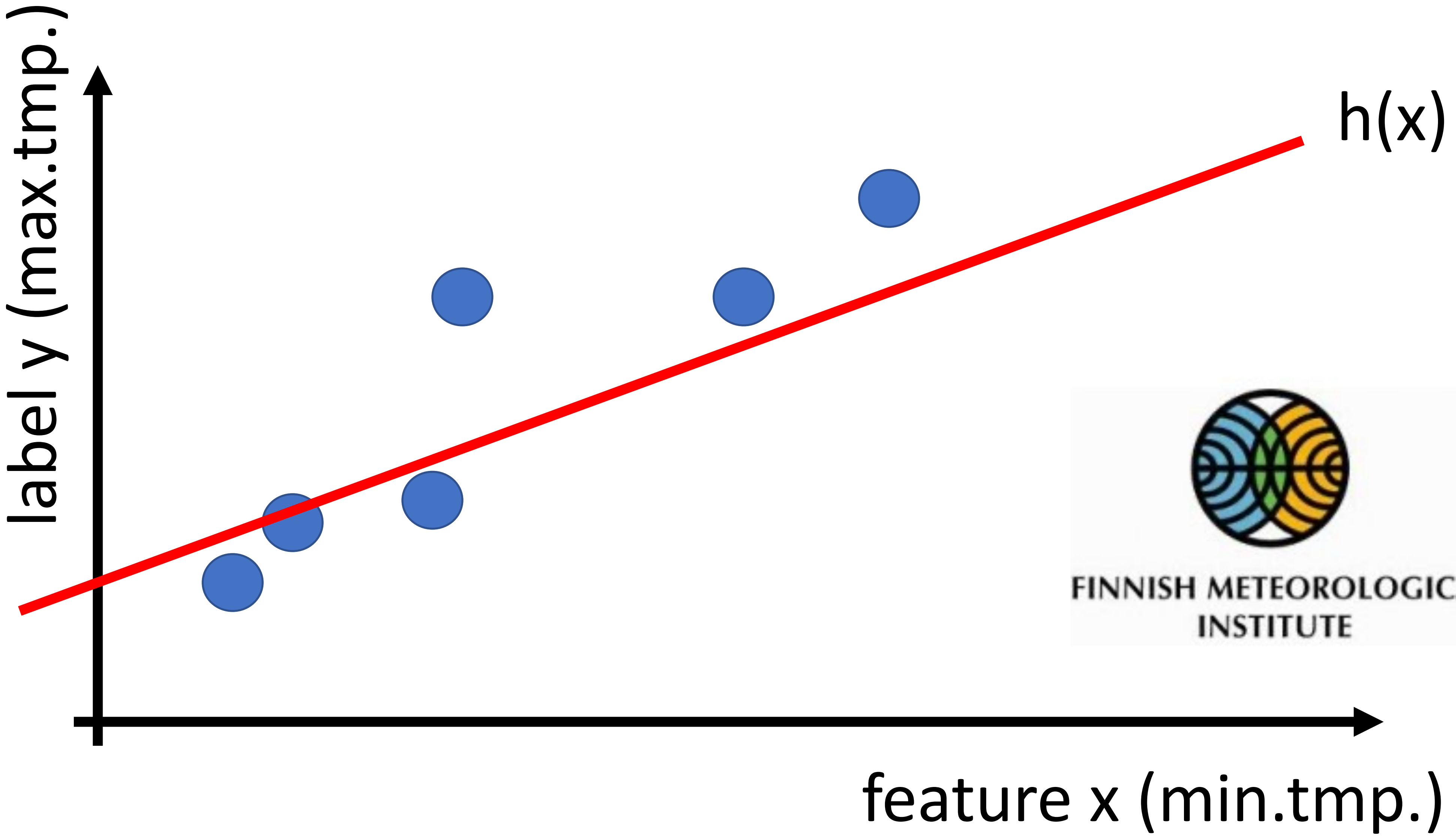
- squared error
- logistic loss
- focal loss
- ….



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

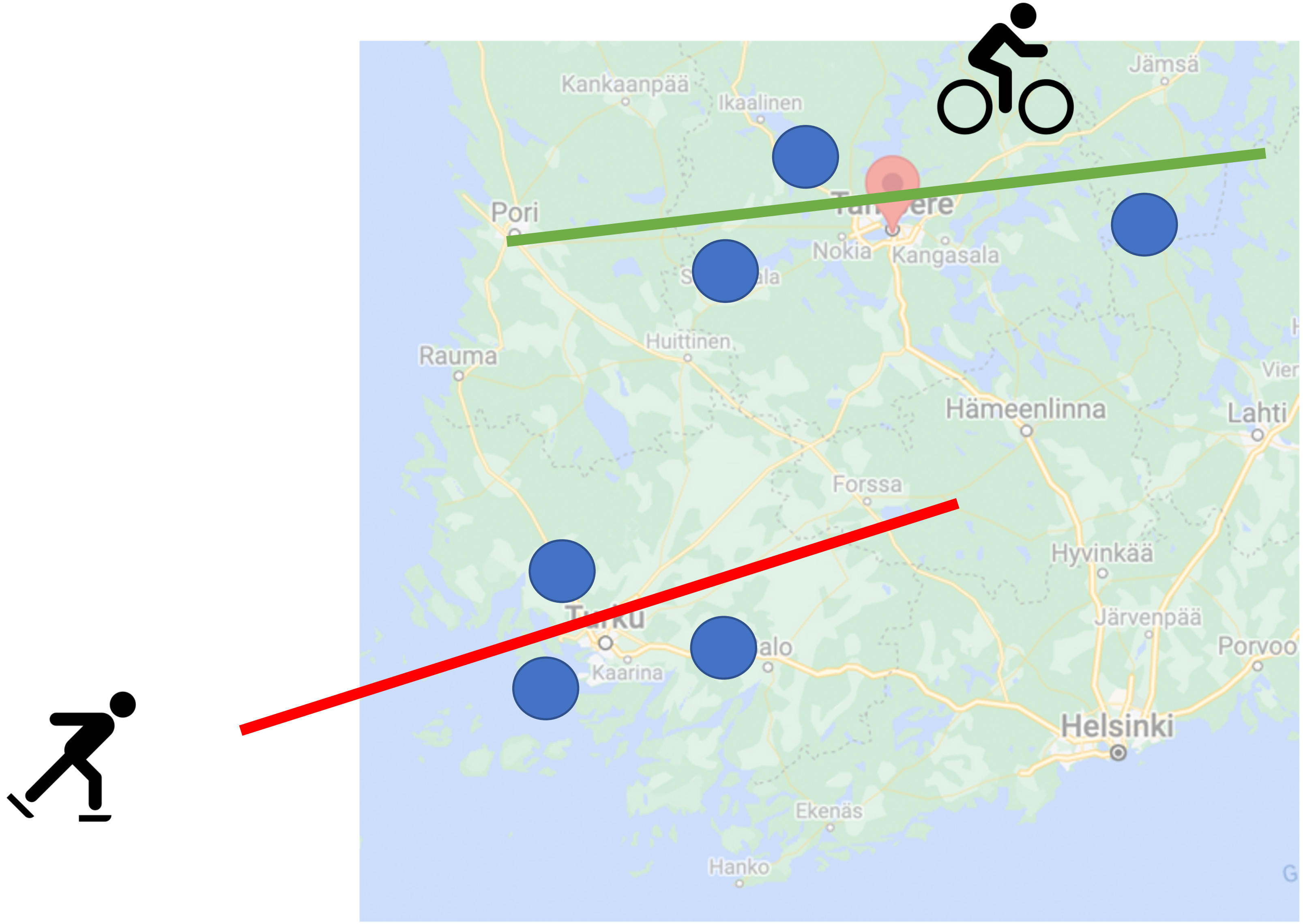
- numeric arrays
- text
- networks
- …

Plain Old Machine Learning.



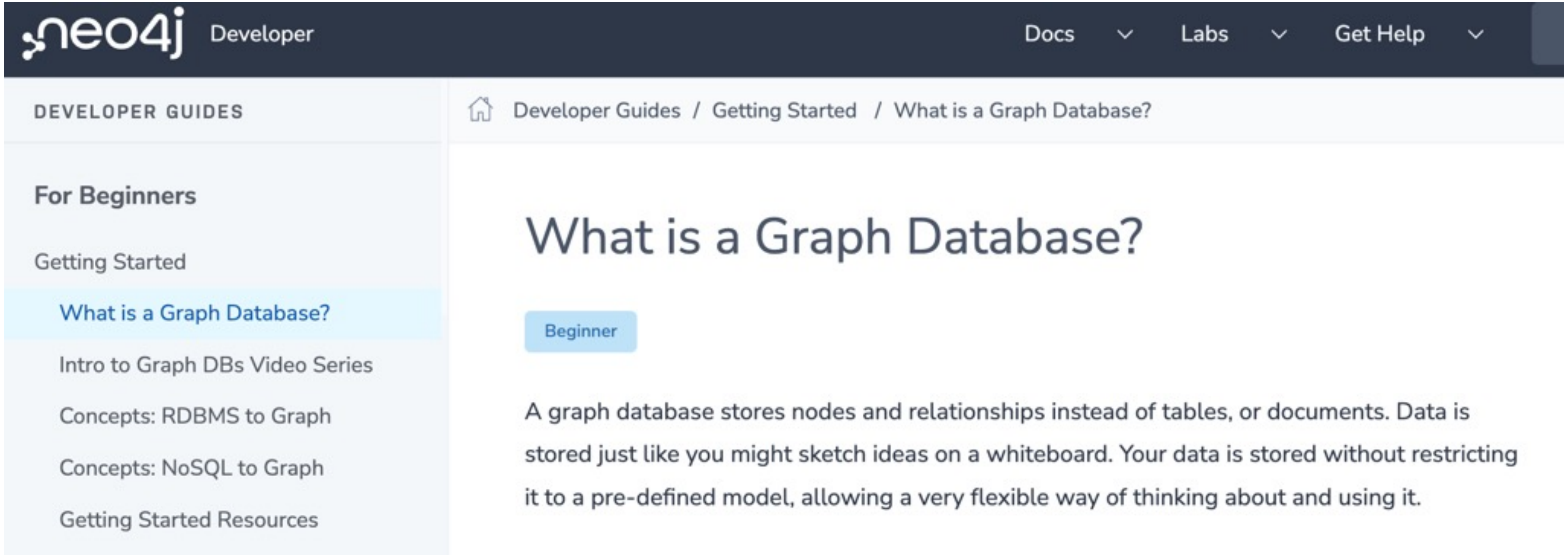
FINNISH METEOROLOGICAL INSTITUTE

Networked Federated Learning



Networked Data

Networked Data=Graph Database



neo4j Developer

Docs Labs Get Help

DEVELOPER GUIDES

Developer Guides / Getting Started / What is a Graph Database?

For Beginners

Getting Started

What is a Graph Database?

Intro to Graph DBs Video Series

Concepts: RDBMS to Graph

Concepts: NoSQL to Graph

Getting Started Resources

What is a Graph Database?

Beginner

A graph database stores nodes and relationships instead of tables, or documents. Data is stored just like you might sketch ideas on a whiteboard. Your data is stored without restricting it to a pre-defined model, allowing a very flexible way of thinking about and using it.

https://neo4j.com/developer/graph-database/

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…"

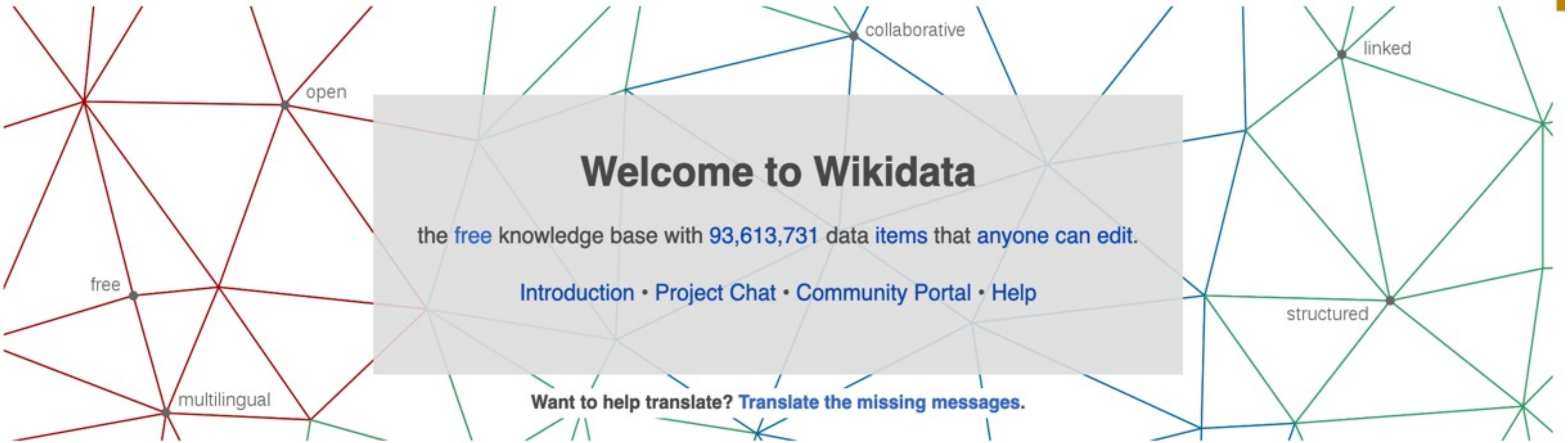
https://image-net.org/

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….."

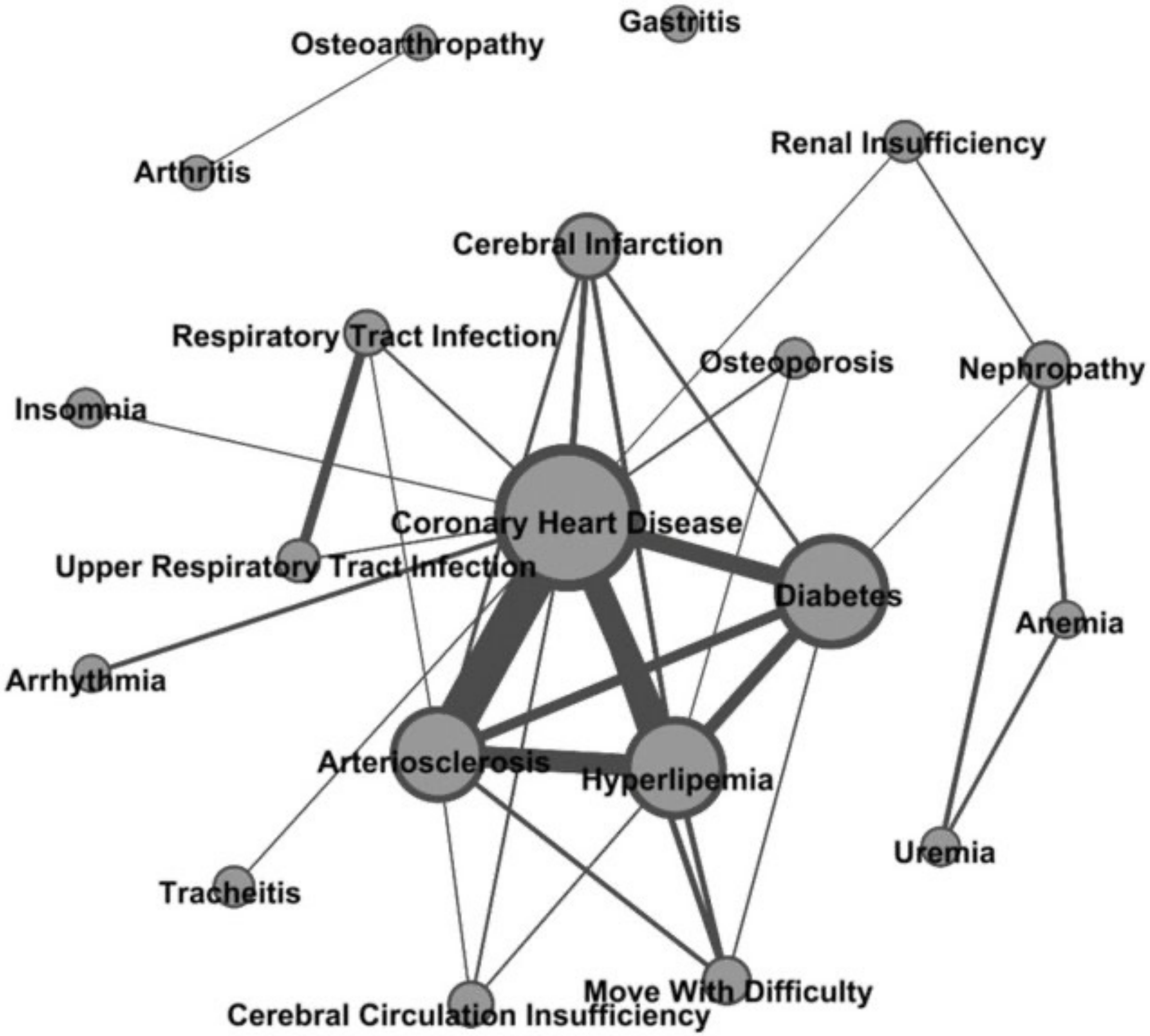
https://wordnet.princeton.edu/

Wikidata.



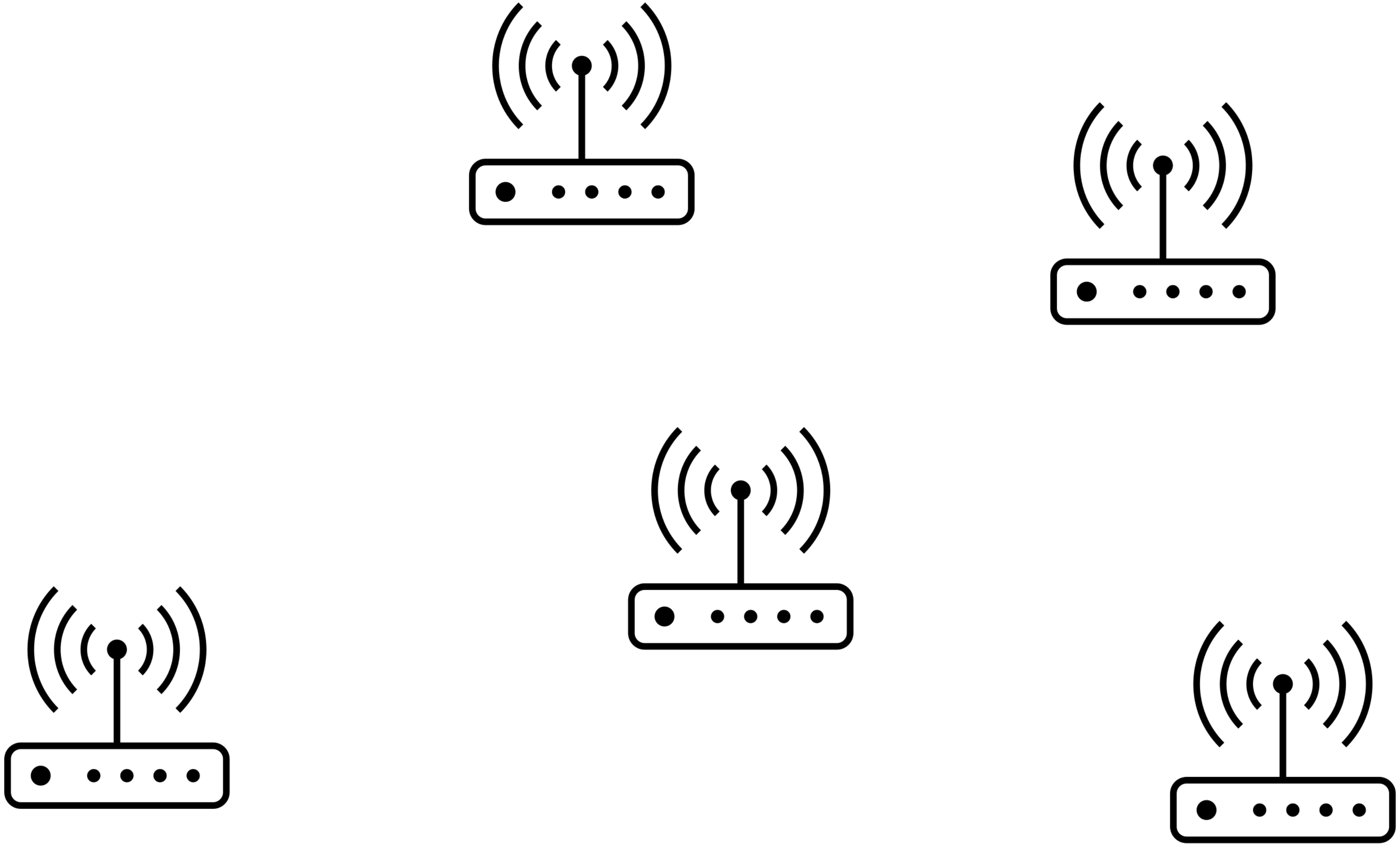
https://www.wikidata.org/wiki/Wikidata:Main_Page

Diseases.



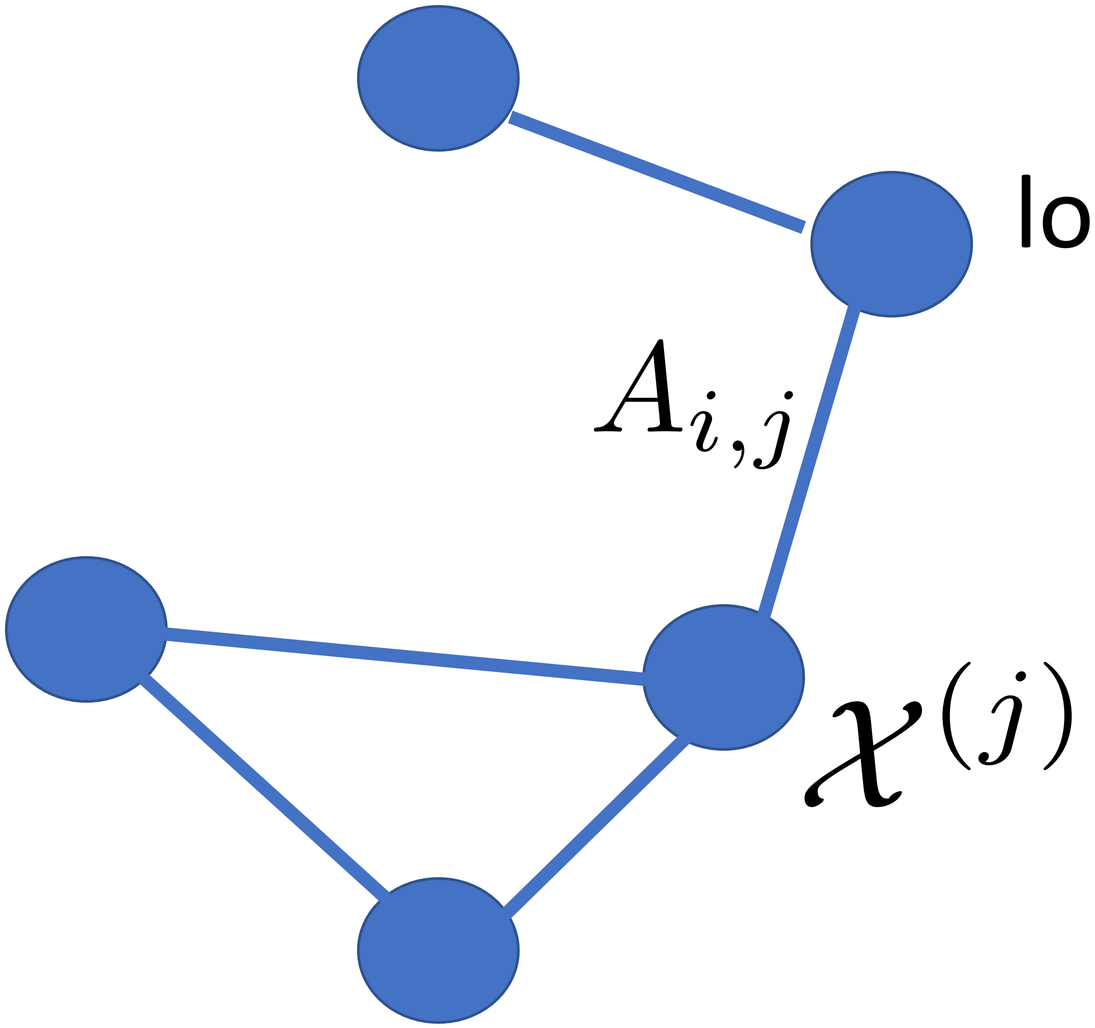
Liu, Jiaqi et.al..
Comorbidity Analysis According to Sex and Age in Hypertension Patients in China.
International Journal of Medical Sciences. 13. 99-107. 10.7150/ijms.13456.

WSN.



Anchors.

Abstraction – The Empirical Graph.



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

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

How To Measure Statistical Sim.?

```
>>> from scipy.stats import ks_2samp
>>> import numpy as np
>>>
>>> np.random.seed(12345678)
>>> x = np.random.normal(0, 1, 1000)
>>> y = np.random.normal(0, 1, 1000)
>>> z = np.random.normal(1.1, 0.9, 1000)
>>>
>>> ks_2samp(x, y)
Ks_2sampResult(statistic=0.022999999999999909, pvalue=0.95189016804849647)
>>> ks_2samp(x, z)
Ks_2sampResult(statistic=0.41800000000000004, pvalue=3.708149411924217e-77)
```

https://stackoverflow.com/questions/10884668/two-sample-kolmogorov-smirnov-test-in-python-scipy

https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

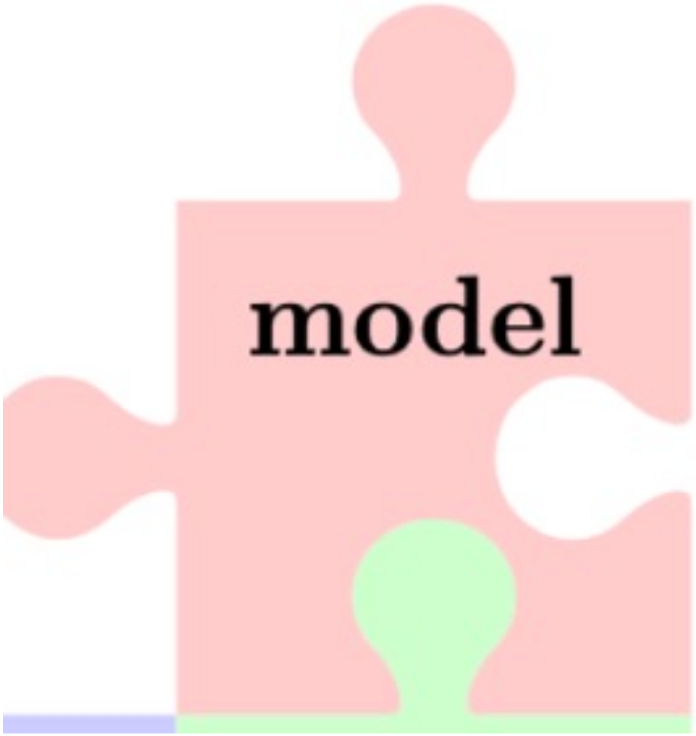
Geometric Dataset Distances via Optimal Transport

David Alvarez-Melis [1] Nicolò Fusi [1]

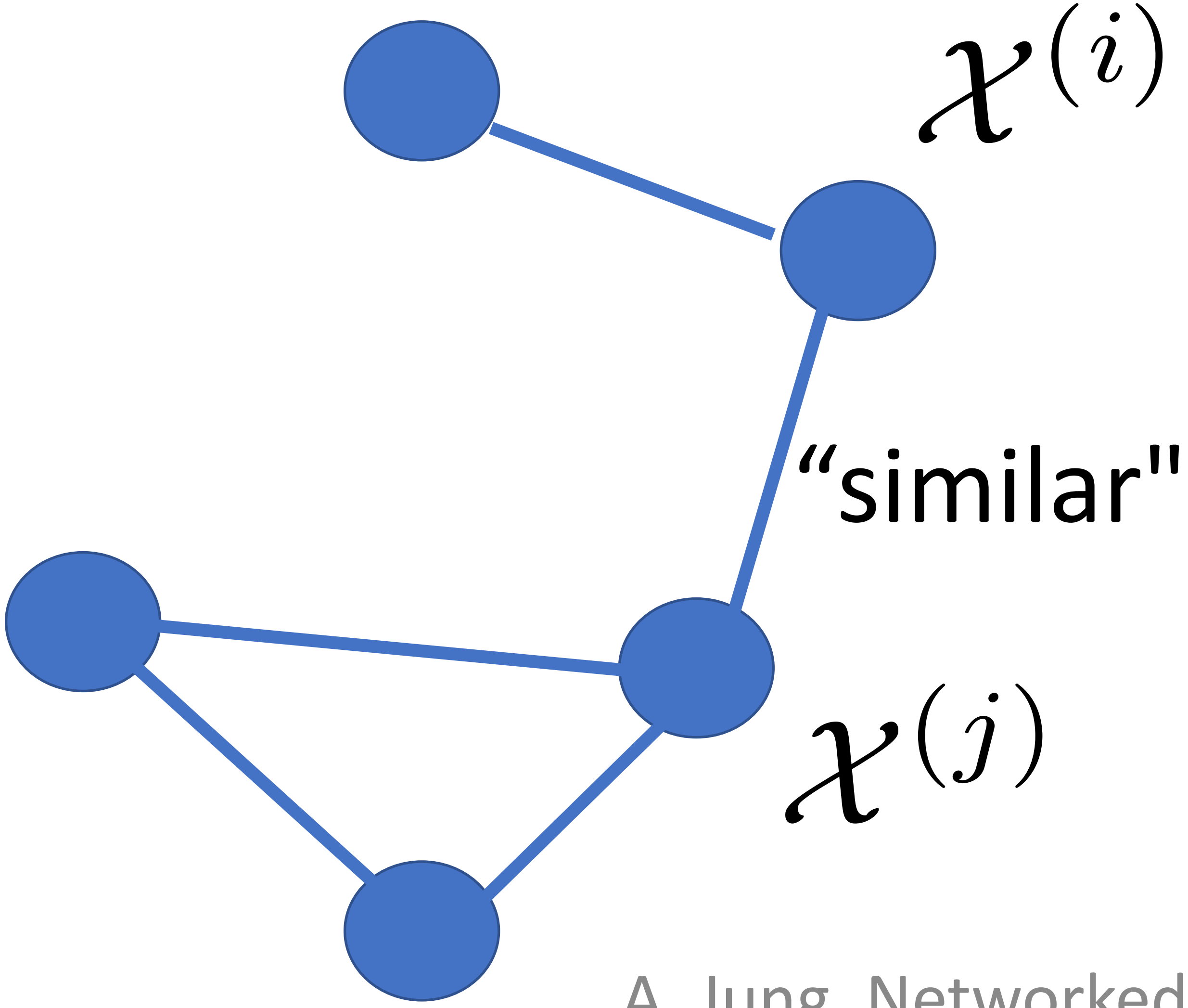
"In this work we propose an alternative notion of distance between datasets that (i) is model-agnostic, (ii) does not involve training,…

https://arxiv.org/pdf/2002.02923.pdf

Networked Models



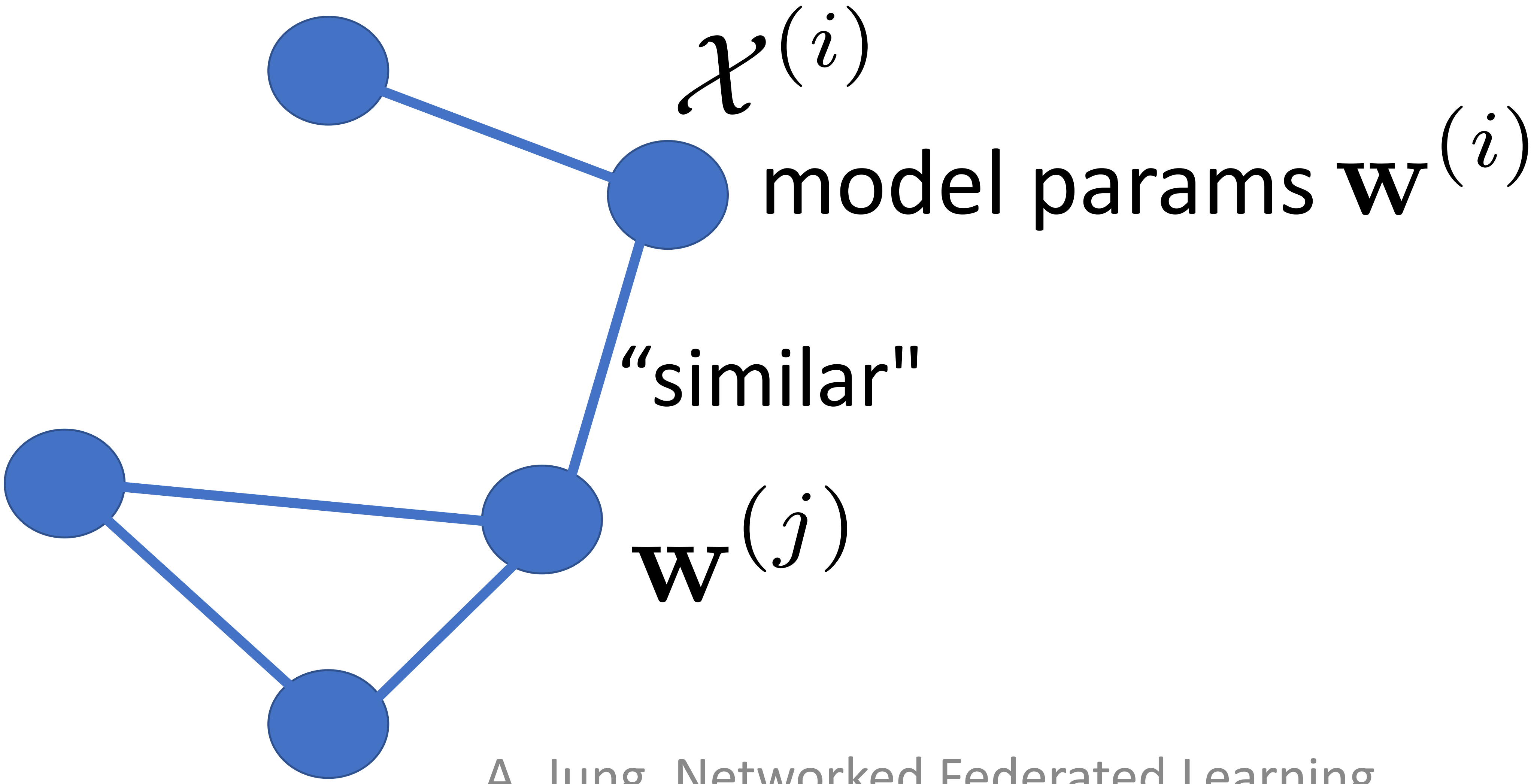
Networked Models.



local model for each node

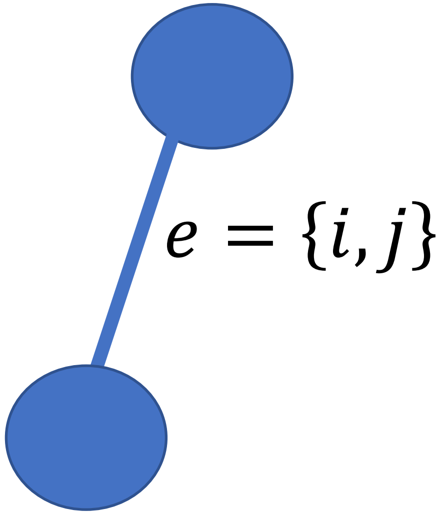
couple models at
connected nodes

Networked Parametric Models.



Smoothness/Clustering Assumption.

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

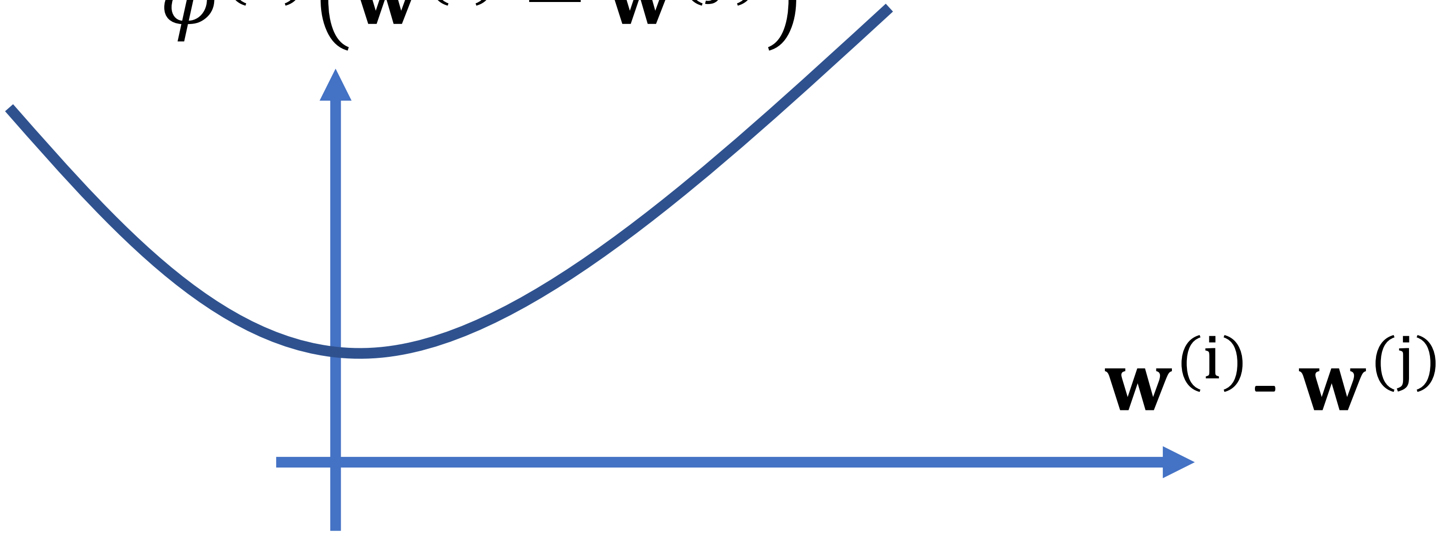


$\boldsymbol{w}^{(j)}$

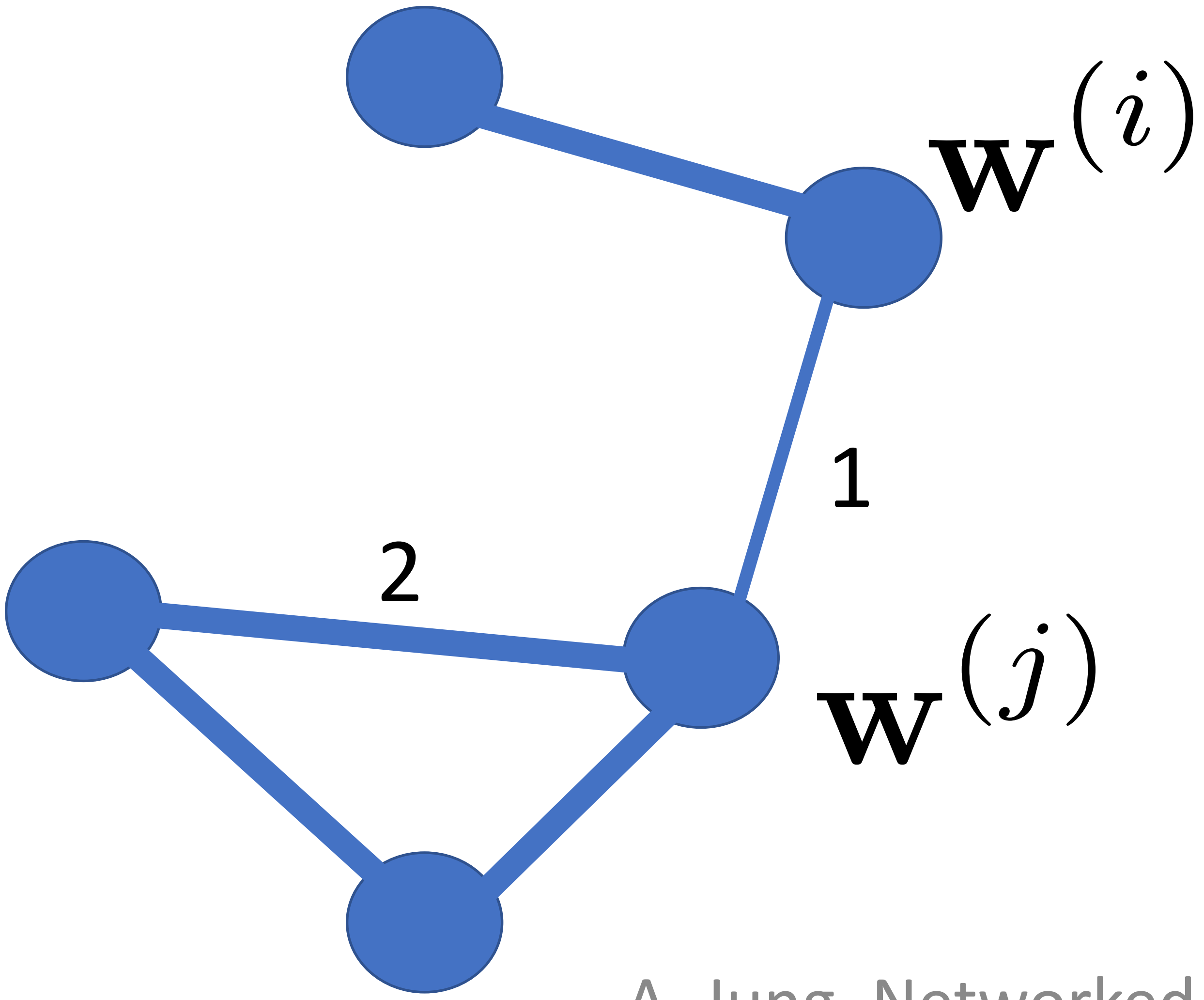
require similar params at ends of edge e

penalty function measures “tension”

$\phi^{(e)}\left(\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\right)$



Generalized Total Variation (GTV)



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

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

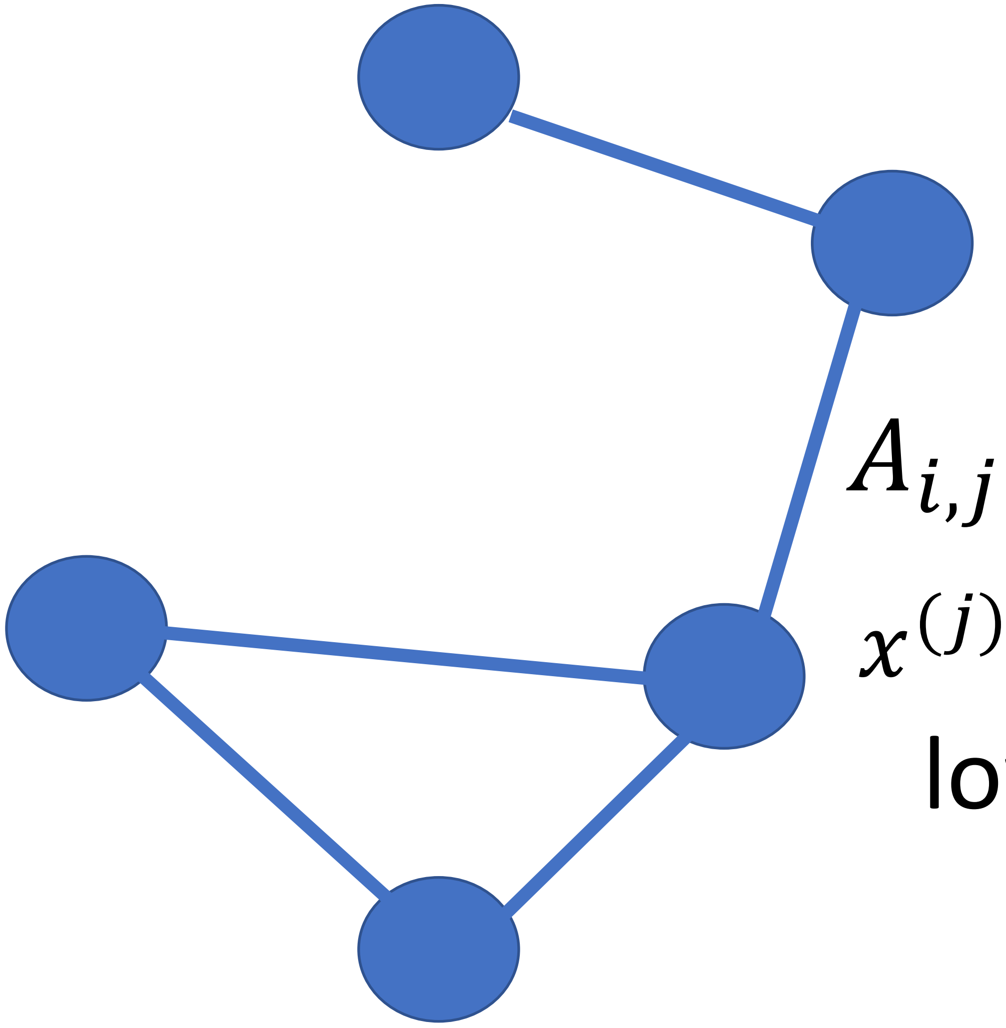
Two Special Cases of GTV.

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.



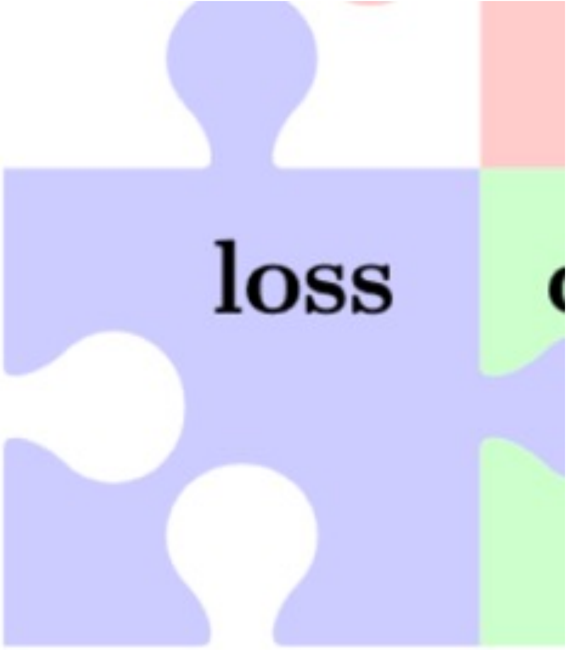
$$x^{(i)} = w^{(i)} + n^{(i)}$$

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

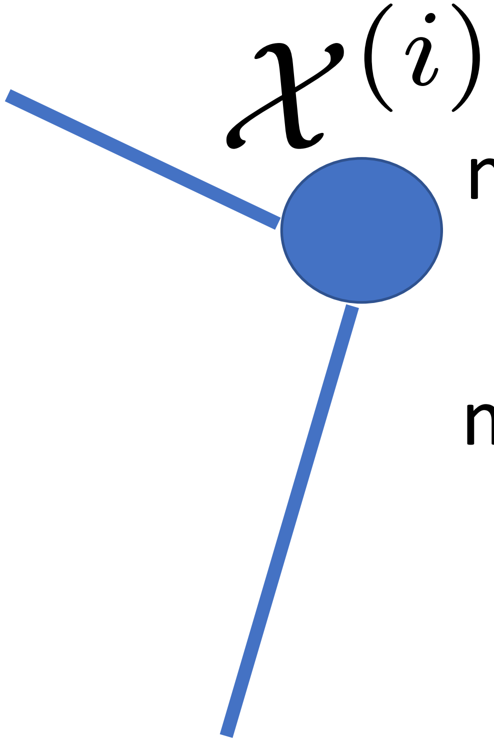
low pass constraint on

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

GTV Minimization.

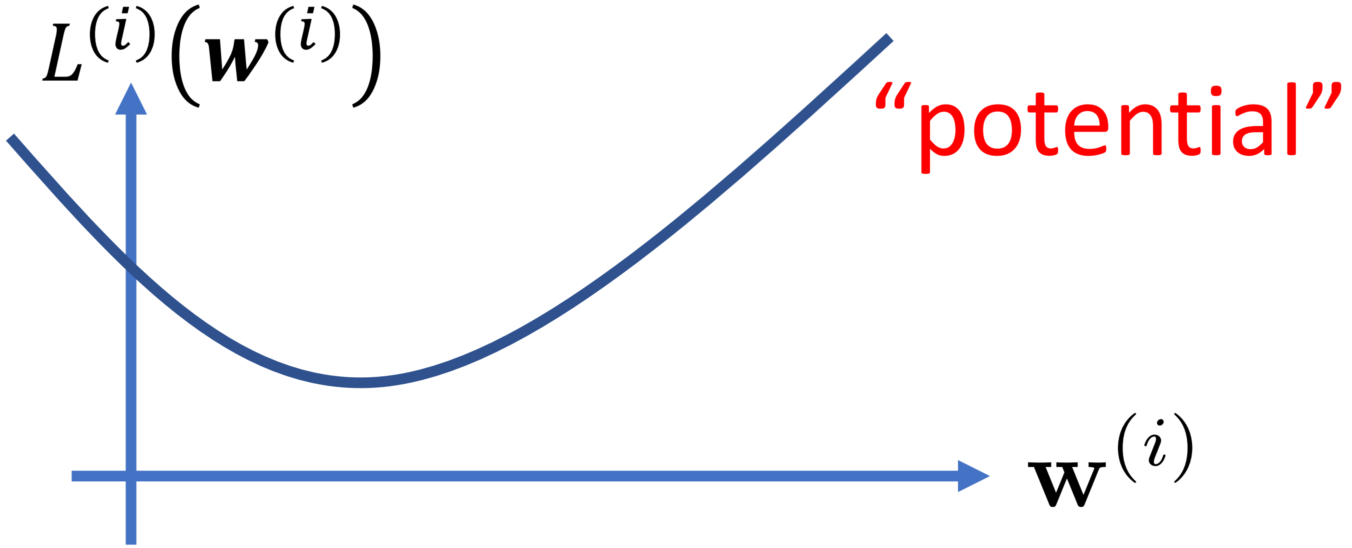


Local Loss Functions.



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

measure quality of params by local loss function



GTV Minimization.

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

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.

- GTVMin as NFL Principle
- The Dual of GTVMin
- Interpretations
- Computational Aspects
- Statistical Aspects

“Massaging” GTV Minimization.

$$\widehat{\mathbf{w}} \in \underset{\mathbf{w} \in \mathcal{W}}{\arg\min} \; f(\mathbf{w}) + g(\mathbf{D}\mathbf{w})$$

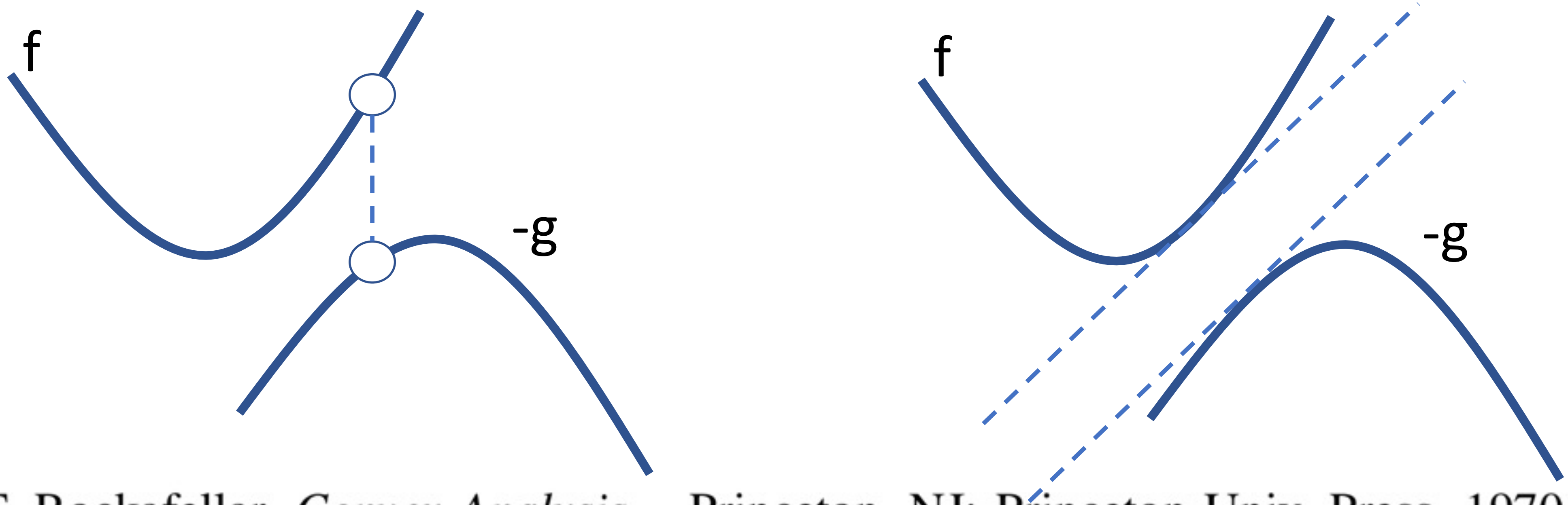
$$\text{with } f(\mathbf{w}) := \sum_{i \in \mathcal{V}} L_i\left(\mathbf{w}^{(i)}\right) \text{ , and } g(\mathbf{u}) := \lambda \sum_{e \in \mathcal{E}} A_e \phi\left(\mathbf{u}^{(e)}\right).$$

with incidence matrix/operator

$$\mathbf{D} : \mathcal{W} \rightarrow \mathcal{U} : \mathbf{w} \mapsto \mathbf{u} \text{ with } \mathbf{u}^{(e)} = \mathbf{w}^{(e_+)} - \mathbf{w}^{(e_-)}.$$

Fenchel's Duality

$$\min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w}) + g(\mathbf{D}\mathbf{w}) = \max_{\mathbf{u}\in\mathcal{U}} -g^{*}(\mathbf{u}) - f^{*}(-\mathbf{D}^{T}\mathbf{u}).$$

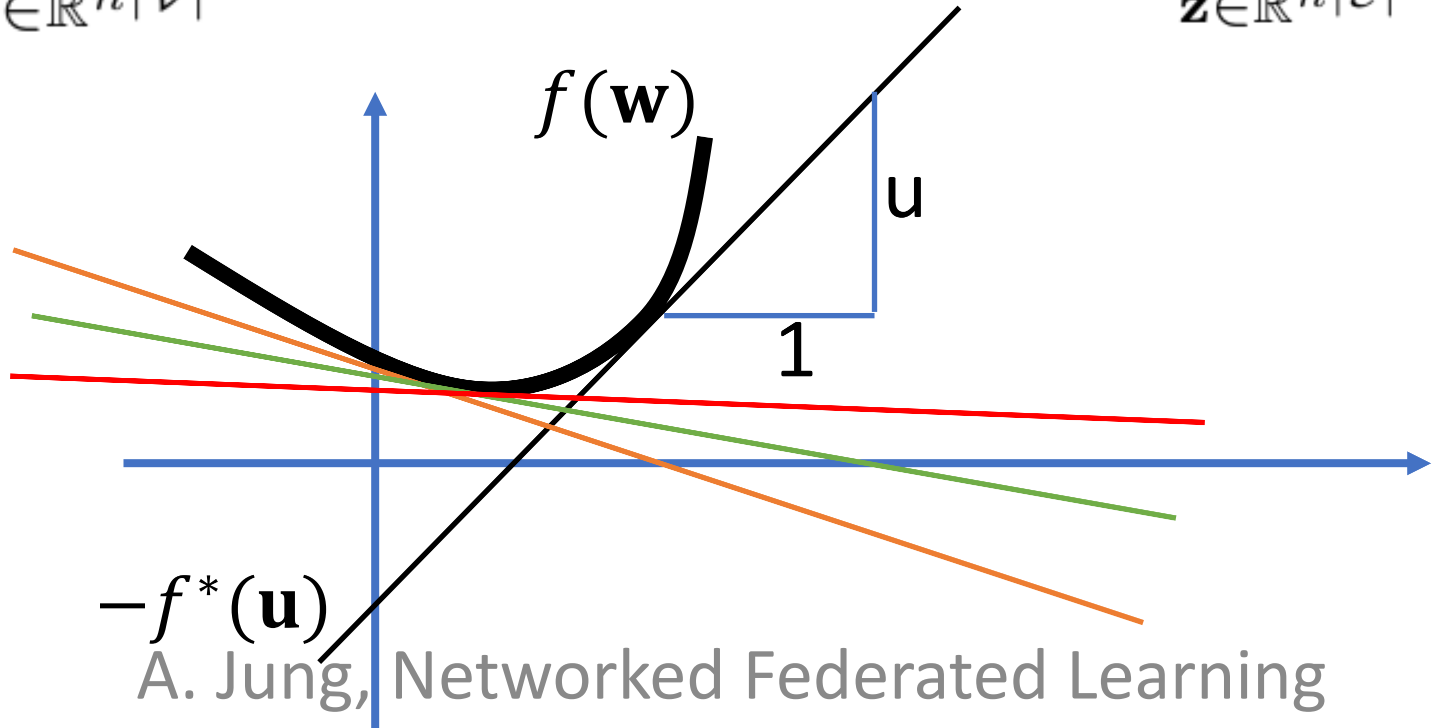


R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.
https://en.wikipedia.org/wiki/Fenchel%27s_duality_theorem

Dual of GTVMin.

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

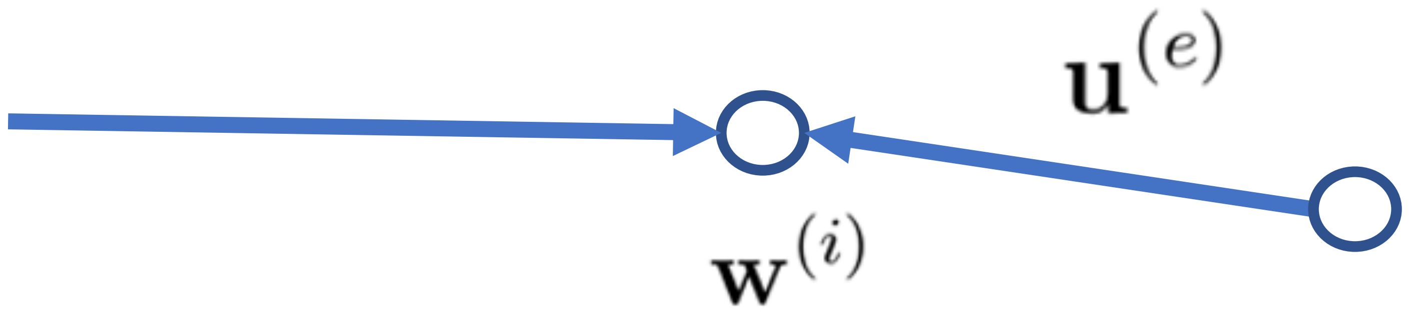
$$f^*(\mathbf{w}) := \sup_{\mathbf{z} \in \mathbb{R}^{n|\mathcal{V}|}} \mathbf{w}^T \mathbf{z} - f(\mathbf{z}) \qquad g^*(\mathbf{u}) := \sup_{\mathbf{z} \in \mathbb{R}^{n|\mathcal{E}|}} \mathbf{u}^T \mathbf{z} - g(\mathbf{z})$$



The Dual of GTVMin.

$$\max_{\mathbf{u} \in \mathcal{U}} - \sum_{i \in \mathcal{V}} L_i^* \left(\mathbf{w}^{(i)}\right) - \lambda \sum_{e \in \mathcal{E}} A_e \phi^* \left(\mathbf{u}^{(e)} / (\lambda A_e)\right)$$

$$\text{subject to } \; -\mathbf{w}^{(i)} = \sum_{e \in \mathcal{E}} \sum_{i = e_{+}} \mathbf{u}^{(e)} - \sum_{i = e_{-}} \mathbf{u}^{(e)} \text{ for all nodes } i \in \mathcal{V}.$$

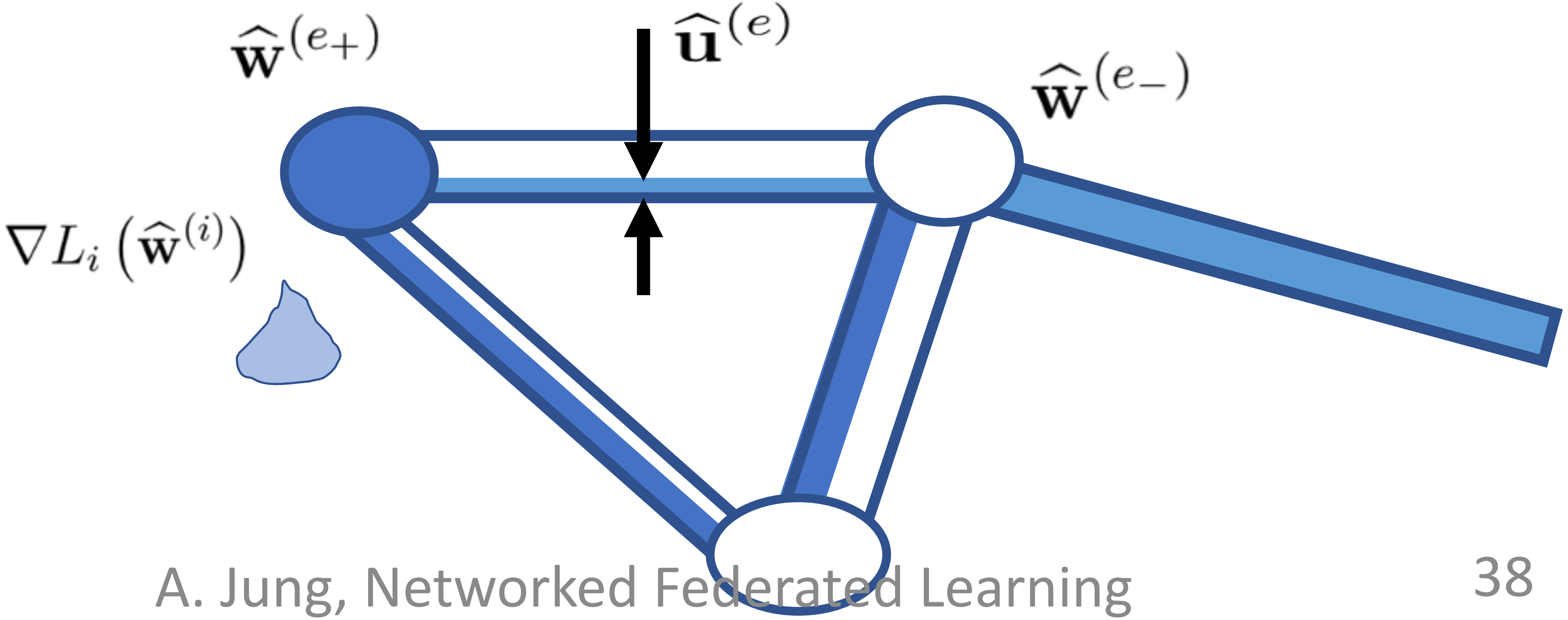


dual variables $\mathbf{u}^{(e)}$ for each (oriented) edge $e = (j, i)$

Primal and Dual Optimality.

$$\sum_{e \in \mathcal{E}} \sum_{i=e_+} \widehat{\mathbf{u}}^{(e)} - \sum_{i=e_-} \widehat{\mathbf{u}}^{(e)} = -\nabla L_i \left(\widehat{\mathbf{w}}^{(i)} \right) \text{ for all nodes } i \in \mathcal{V}$$

$$\widehat{\mathbf{w}}^{(e_+)} - \widehat{\mathbf{w}}^{(e_-)} \in (\lambda A_e) \partial \phi^* (\widehat{\mathbf{u}}^{(e)} / (\lambda A_e)) \text{ for every edge } e \in \mathcal{E}.$$



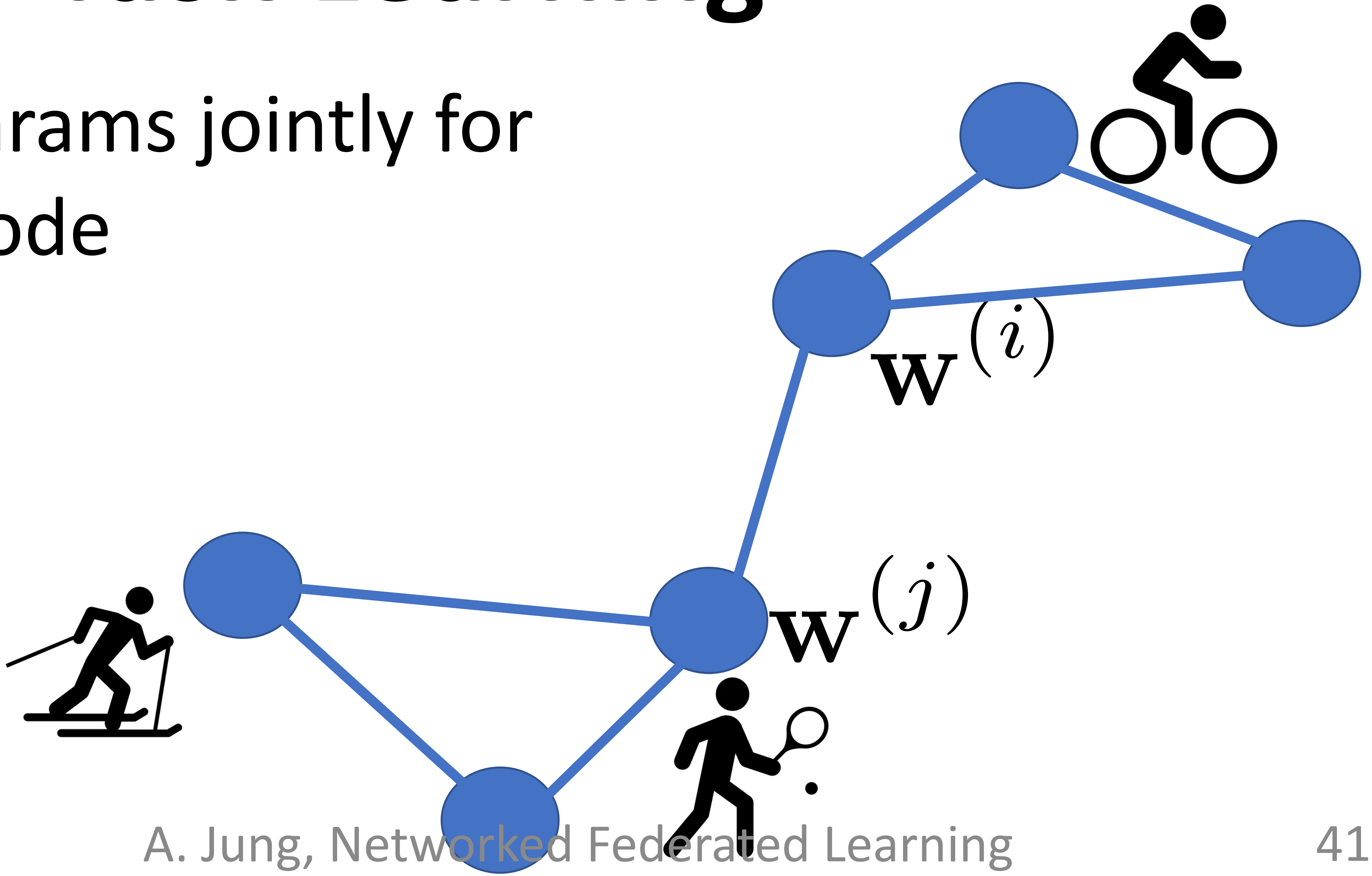
- GTVMin as NFL Principle
- The Dual of GTVMin
- Interpretations
- Computational Aspects
- Statistical Aspects

Smooth Graph Sig. Recovery

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

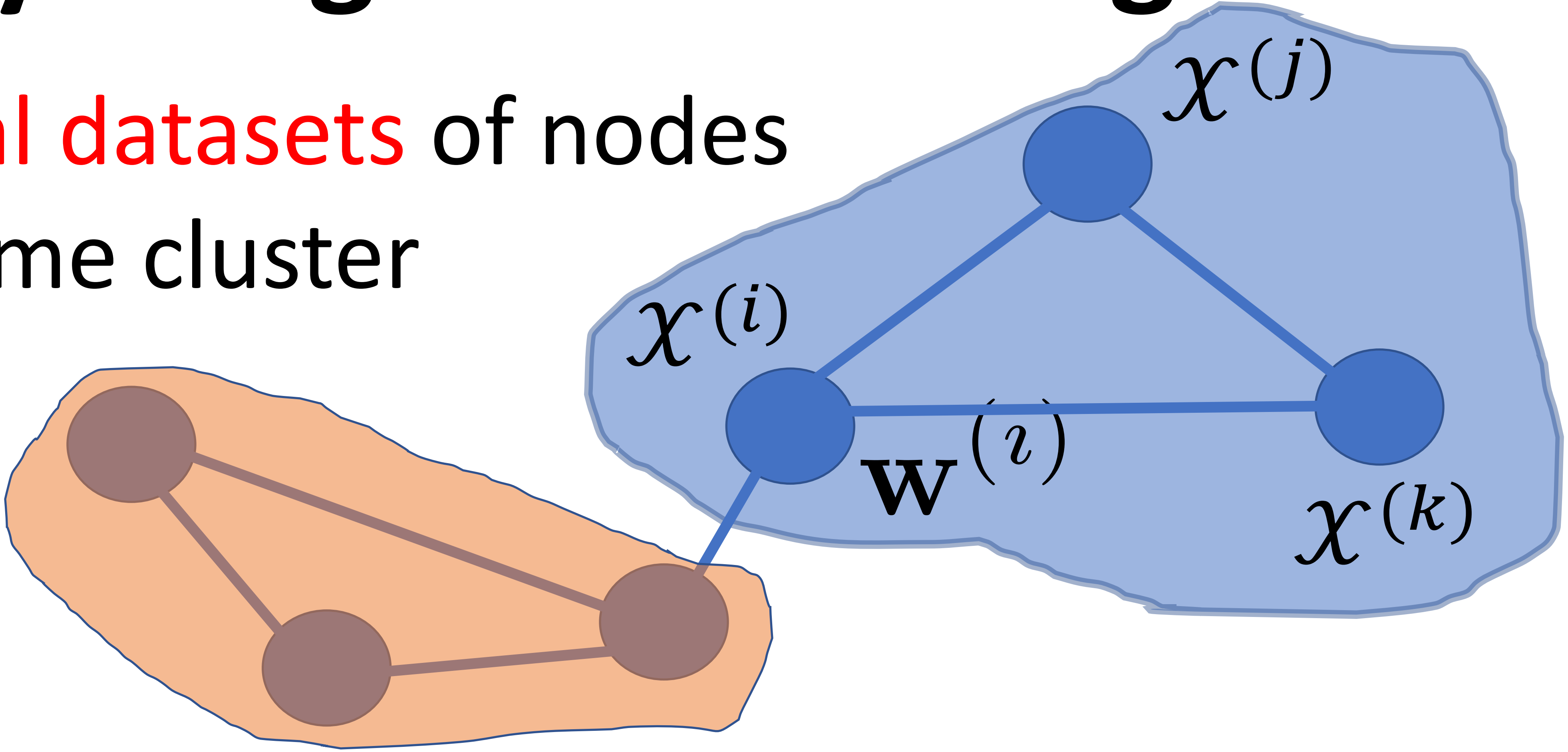
Multi-Task Learning

learn params jointly for every node



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.

Generalized Convex Clustering

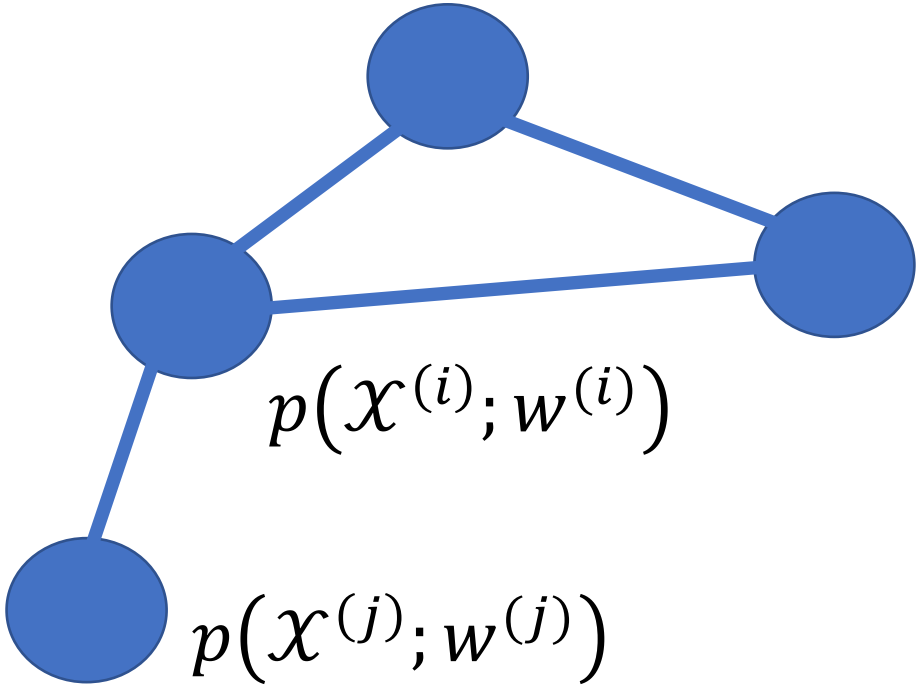
$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} \left\| w^{(i)} - a^{(i)} \right\|^2 + \lambda \sum_{\{i,j\}} A_{i,j} \left\| w^{(i)} - w^{(j)} \right\|_p$$

D. Sun, K.-C. Toh, Y. Yuan;
Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm, JMLR, 22(9):1–32, 2021

(Probabilistic) Graphical Model

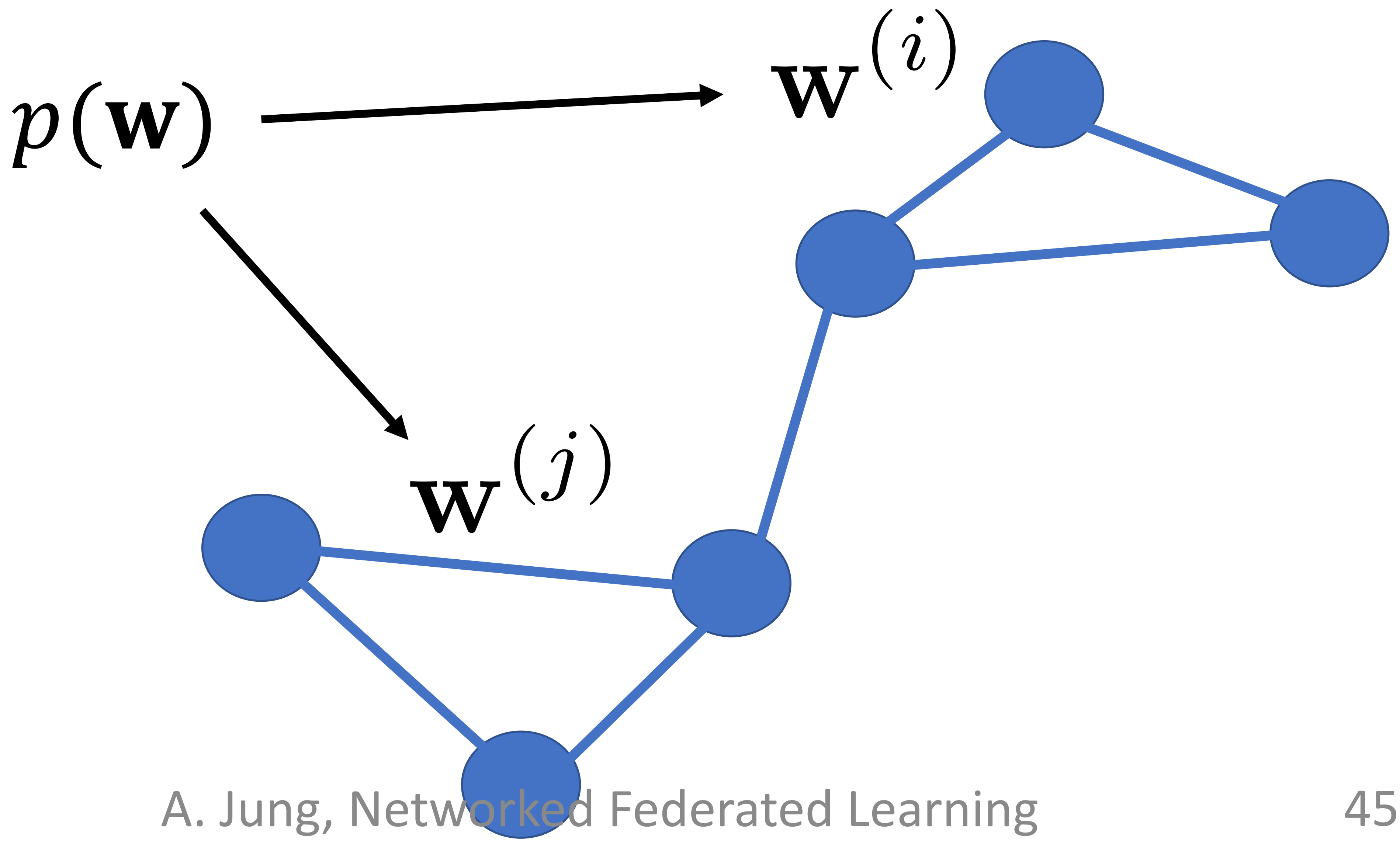
separate prob. space for each local dataset

traditionally, PGMs use a common prob. space for all local datasets



AJ, "**Networked Exponential Families for Big Data Over Networks,**" in *IEEE Access*, vol. 8, pp. 202897-202909, 2020, doi: 10.1109/ACCESS.2020.3033817.

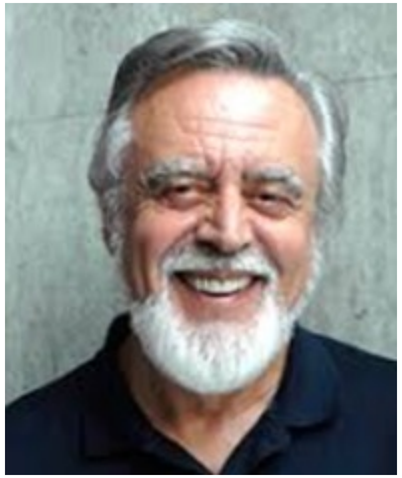
Approx. Hierarch. Bayes' Model

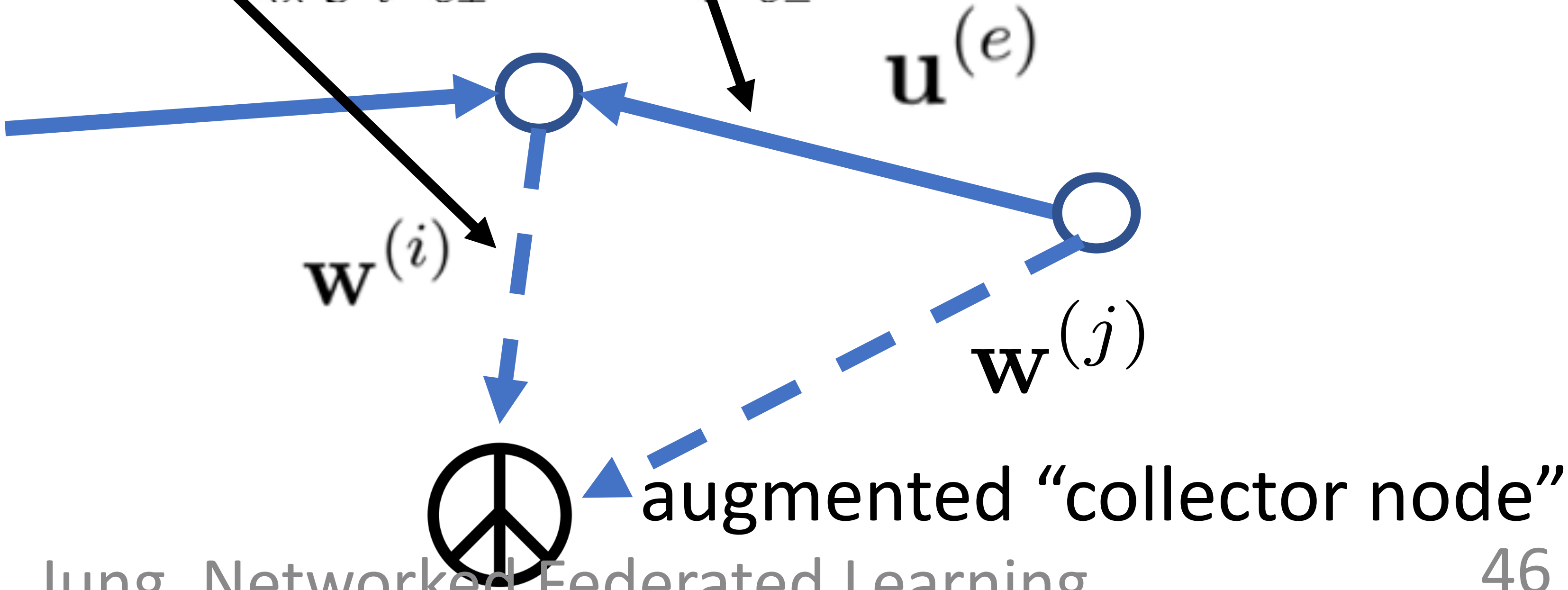


Vector-Valued Min-Cost-Flow

$$\max_{\mathbf{u}\in\mathcal{U}} -\sum_{i\in\mathcal{V}} L_i^*\left(\mathbf{w}^{(i)}\right) - \lambda \sum_{e\in\mathcal{E}} A_e \phi^*\left(\mathbf{u}^{(e)}/(\lambda A_e)\right)$$

$$\text{subject to } \; -\mathbf{w}^{(i)} = \sum_{e\in\mathcal{E}} \sum_{i=e_+} \mathbf{u}^{(e)} - \sum_{i=e_-} \mathbf{u}^{(e)} \text{ for all nodes } i \in \mathcal{V}.$$





Electrical Network.

Kirchhoff's Current Law

$$\sum_{e \in \mathcal{E}} \sum_{i=e_{+}} \widehat{\mathbf{u}}^{(e)} - \sum_{i=e_{-}} \widehat{\mathbf{u}}^{(e)} = -\nabla L_i \left(\widehat{\mathbf{w}}^{(i)} \right) \text{ for all nodes } i \in \mathcal{V}$$

$$\widehat{\mathbf{w}}^{(e_{+})} - \widehat{\mathbf{w}}^{(e_{-})} \in (\lambda A_e) \partial \phi^{*} (\widehat{\mathbf{u}}^{(e)} / (\lambda A_e)) \text{ for every edge } e \in \mathcal{E}.$$

Generalized Ohm Law

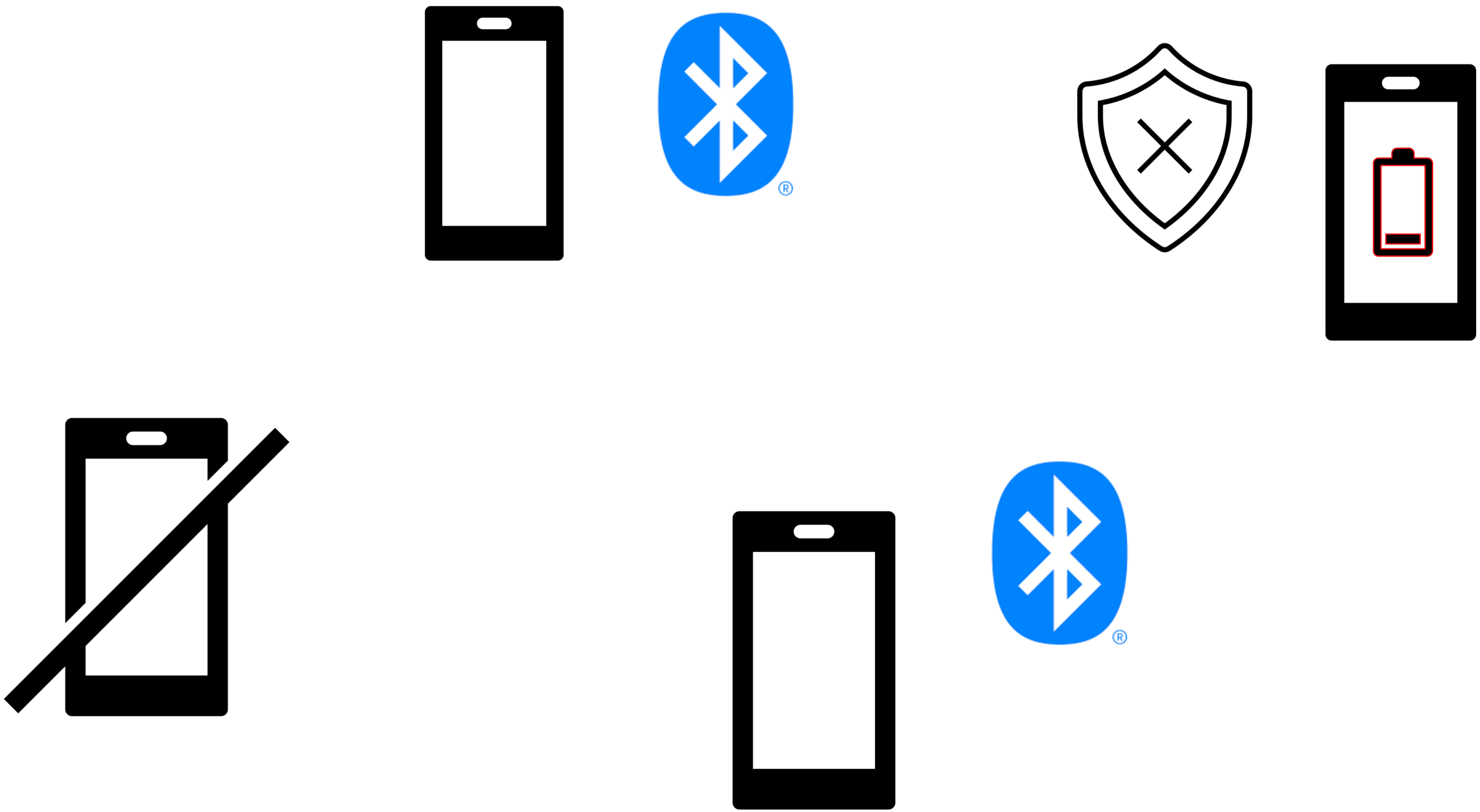
- GTVMin as NFL Principle
- The Dual of GTVMin
- Interpretations
- Computational Aspects
- Statistical Aspects

Computational Aspects.

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

- solve in ad-hoc nets of low-cost devices
- robustness against node/link failures
- robustness against “stragglers”

Our Toy NFL Setting



Another NFL Setting…

https://www.google.com/about/datacenters/

https://en.wikipedia.org/wiki/Optical_fiber

Two Main Flavours

- Primal (Gradient) Methods

- Primal-Dual Methods

Two Main Flavours

- Primal (Gradient) Methods

- Primal-Dual Methods

Gradient Descent

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

optimality condition $\nabla f(w) = 0$

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

Subgradient Descent (SGD)

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

optimality condition $0 \,\epsilon \partial f(w)$

$$w^{(k+1)} = w^{(k)} - \alpha^{(k)} g^{(k)} \qquad g^{(k)} \epsilon \partial f\left(w^{(k)}\right)$$

Distributed SGD

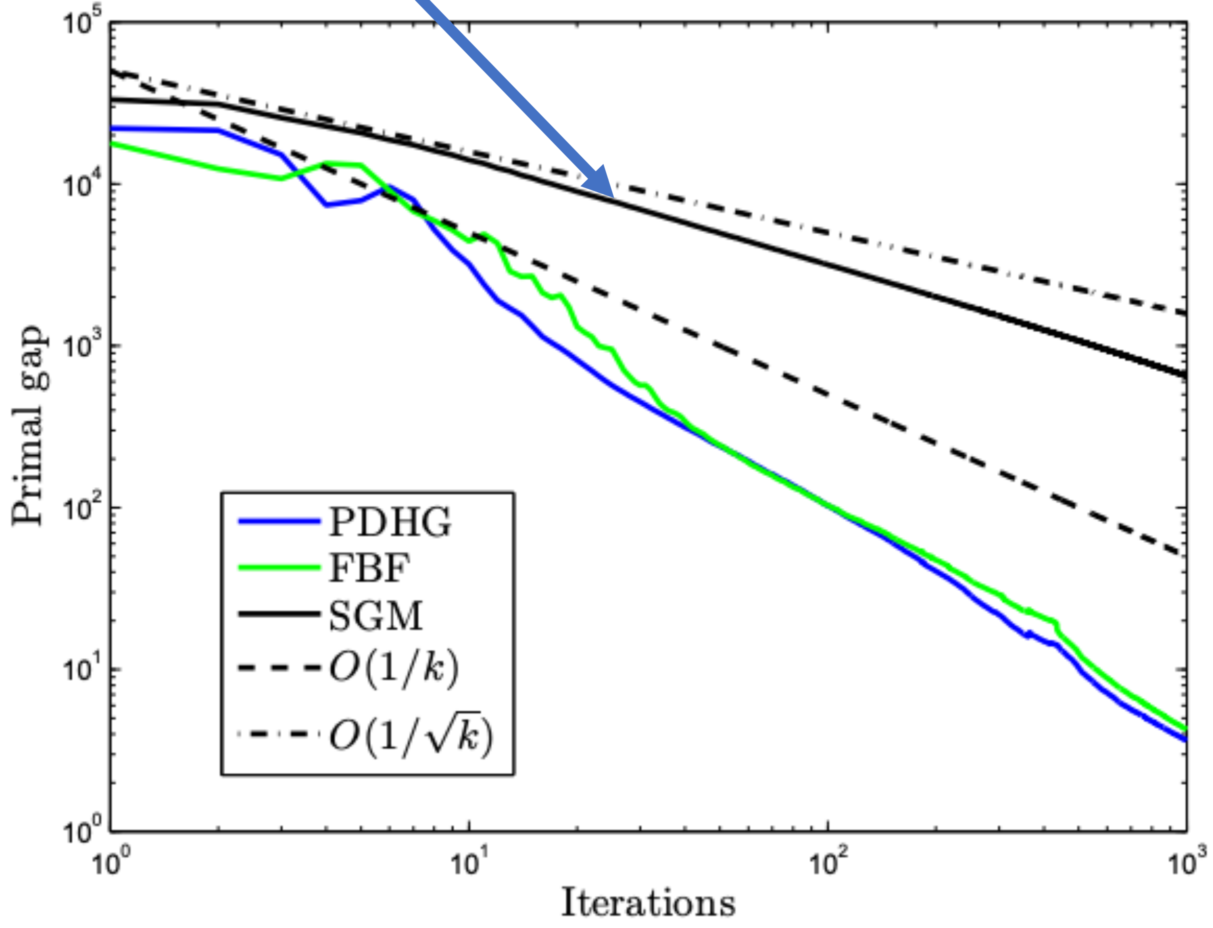
A. Nedić and A. Olshevsky, "Distributed Optimization Over Time-Varying Directed Graphs," in *IEEE Transactions on Automatic Control*, 2015,

A. Nedic (M.S., University of Belgrade, 1991)

A. Nedić and A. Olshevsky, "Stochastic Gradient-Push for Strongly Convex Functions on Time-Varying Directed Graphs," in *IEEE Transactions on Automatic Control*, 2016,

A. Nedic and A. Ozdaglar, "Distributed Subgradient Methods for Multi-Agent Optimization," in *IEEE Transactions on Automatic Control*, Jan. 2009.

SGD Requires Many Iter.



Complexity of SGD

Theorem 3. *Let $L_\ell, R > 0$ and $\gamma \in (0,1]$. There exists a matrix W of eigengap $\gamma(W) = \gamma$, and n functions f_i satisfying (A2), where n is the size of W, such that for all $t < \frac{d-2}{2}\min(\tau/\sqrt{\gamma}, 1)$ and all $i \in \{1, ..., n\}$,*

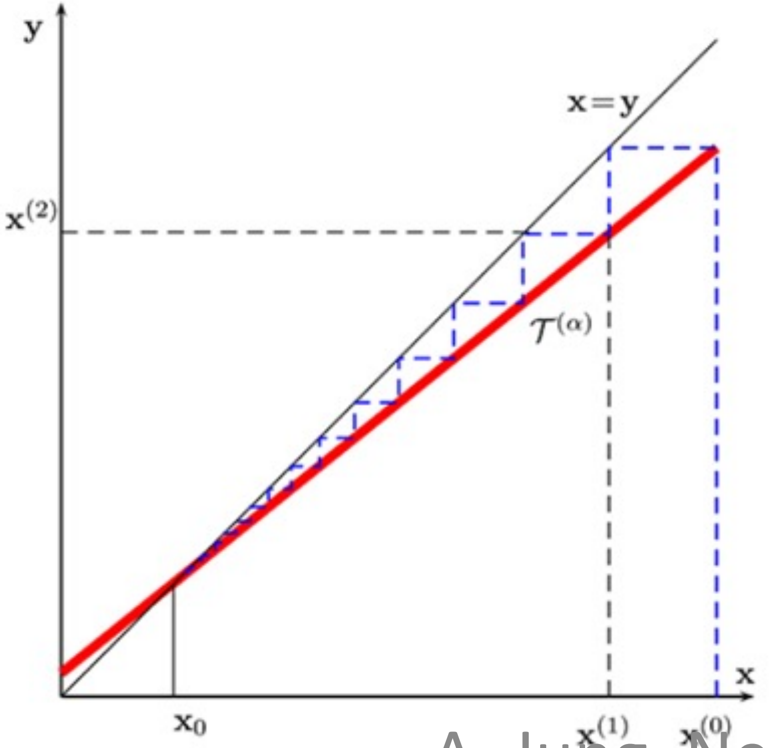
$$\bar{f}(\theta_{i,t}) - \min_{\theta \in B_2(R)} \bar{f}(\theta) \geq \frac{RL_\ell}{108}\sqrt{\frac{1}{(1+\frac{2t\sqrt{\gamma}}{\tau})^2} + \frac{1}{1+t}}. \tag{19}$$

K. Scaman, F. Bach, S. Bubeck, L. Massoulié, Y Lee, Optimal Algorithms for Non-Smooth Distributed Optimization in Networks, NeurIPS 2018.

SGD as Fixed Point Iteration

$$w^{(k+1)} = \mathcal{T}^{(k)}\left(w^{(k)}\right)$$

with $\mathcal{T}^{(k)}\left(w^{(k)}\right) = w^{(k)} - \alpha^{(k)}\partial f\left(w^{(k)}\right)$



AJ, "A Fixed-Point of View on Gradient Methods for Big Data", Front. Appl. Math. Stat., 2017.

Plenary on Fixed-Point Tools

$$w^{(k+1)} = \mathbb{Q}^{(k)}\left(w^{(k)}\right)$$

Jean-Christophe Pesquet

Jean-Christophe Pesquet (I
in 1987, the Ph.D. and HDR
1999, he was a Maître de C
University Paris-Est, and fro
the university. He is curre
Director of the CVN (Inria te
2021. In 2005, J.-C. Pesque
was a member of the SPTM
IEEE SPL (2004-2006). He
journal (2010-2015), and a r
now an associate editor of
methods in data science.

Fixed Point Strategies in Signal and Image Processing

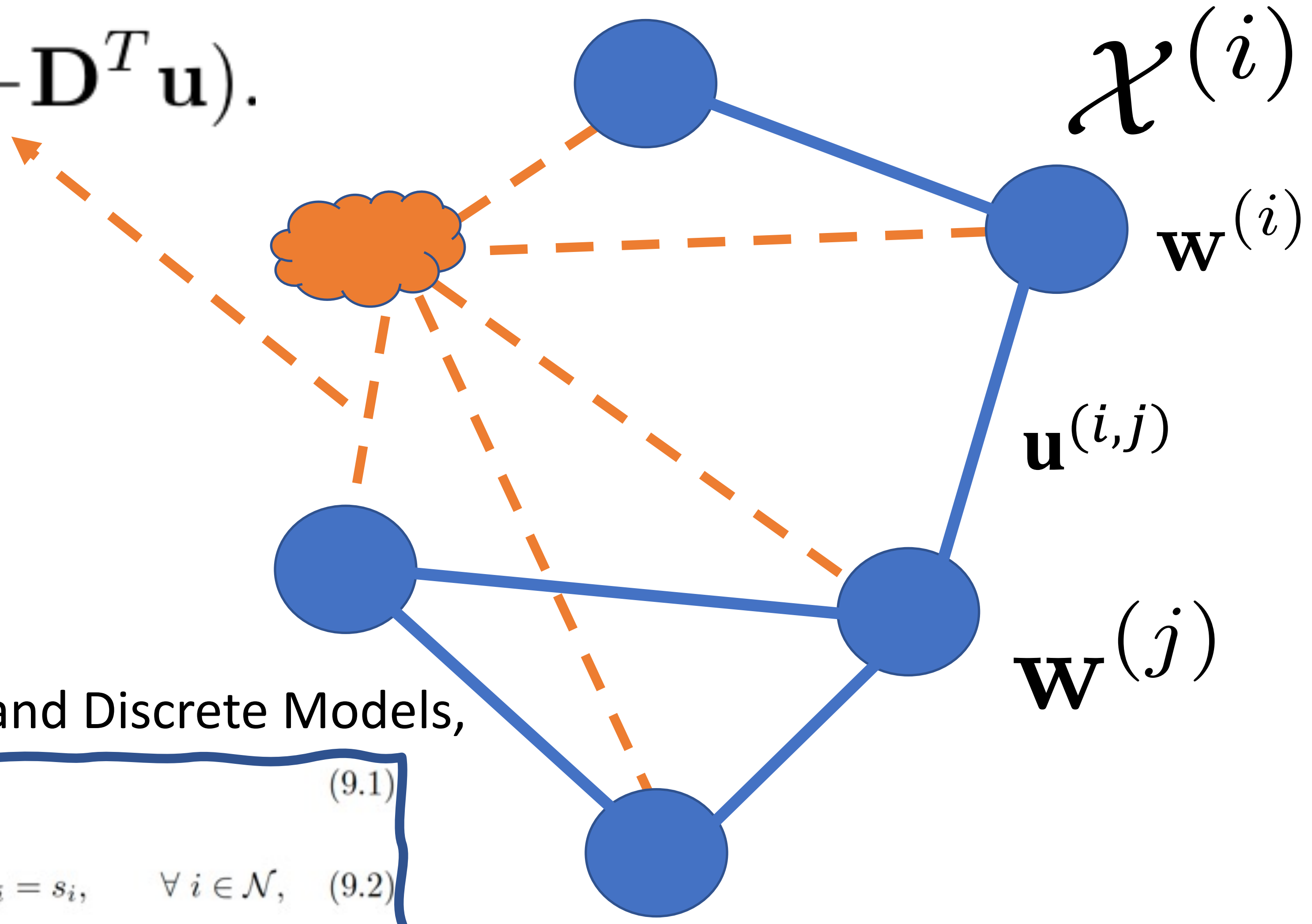
Two Main Flavours

- Primal (Gradient) Methods

- Primal-Dual Methods

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 penalty)

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

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

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

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

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.

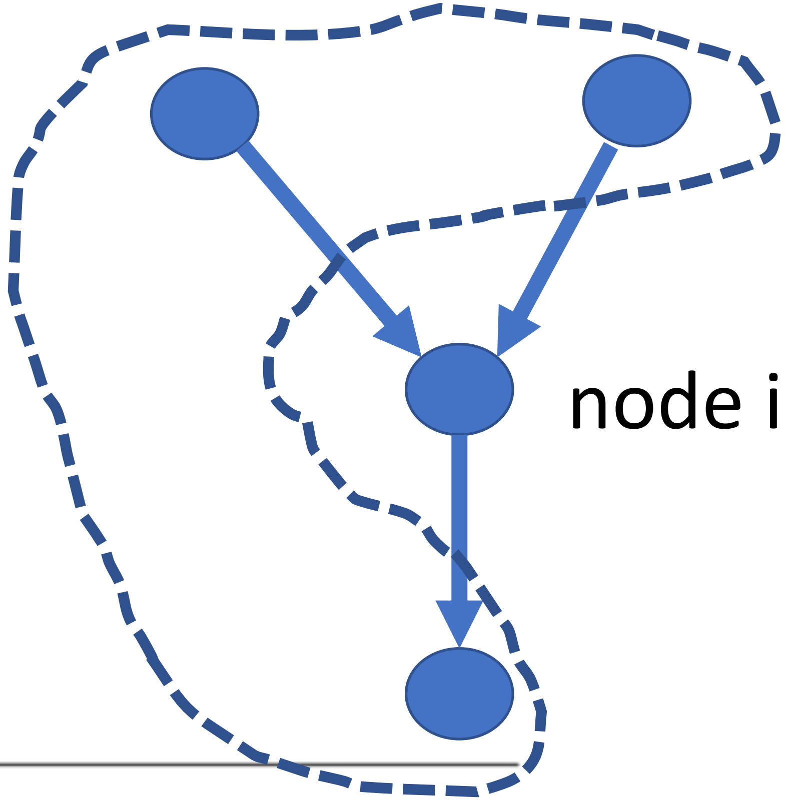
After Some Manipulations.

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**
    $\widehat{\mathbf{w}}_{k+1}^{(i)} := \widehat{\mathbf{w}}_k^{(i)} - \tau_i \sum_{e\in\mathcal{E}} D_{e,i}\widehat{\mathbf{u}}_k^{(e)}$
  end for
  for nodes in the training set $i \in \mathcal{M}$ **do**
    $\widehat{\mathbf{w}}_{k+1}^{(i)} := \mathcal{PU}^{(i)}\{\widehat{\mathbf{w}}_{k+1}^{(i)}\}$
  end for
  for all edges $e \in \mathcal{E}$ **do**
    $\widehat{\mathbf{u}}_{k+1}^{(e)} := \widehat{\mathbf{u}}_k^{(e)} + \sigma_e\big(2\big(\widehat{\mathbf{w}}_{k+1}^{(e_+)} - \widehat{\mathbf{w}}_{k+1}^{(e_-)}\big) - \big(\widehat{\mathbf{w}}_k^{(e_+)} - \widehat{\mathbf{w}}_k^{(e_-)}\big)\big)$
    $\widehat{\mathbf{u}}_{k+1}^{(e)} := \mathcal{DU}^{(e)}\{\widehat{\mathbf{u}}_{k+1}^{(e)}\}$
  end for
  $k := k+1$
end while

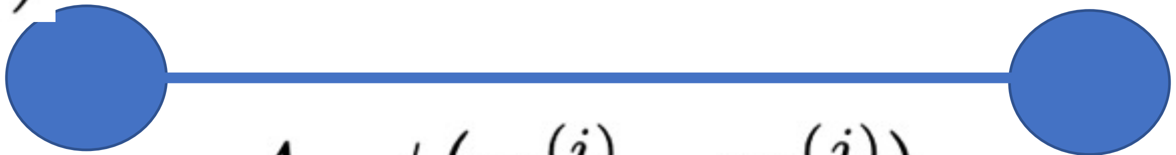


Algorithm 1 is Attractive for NFL…

- 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.

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



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

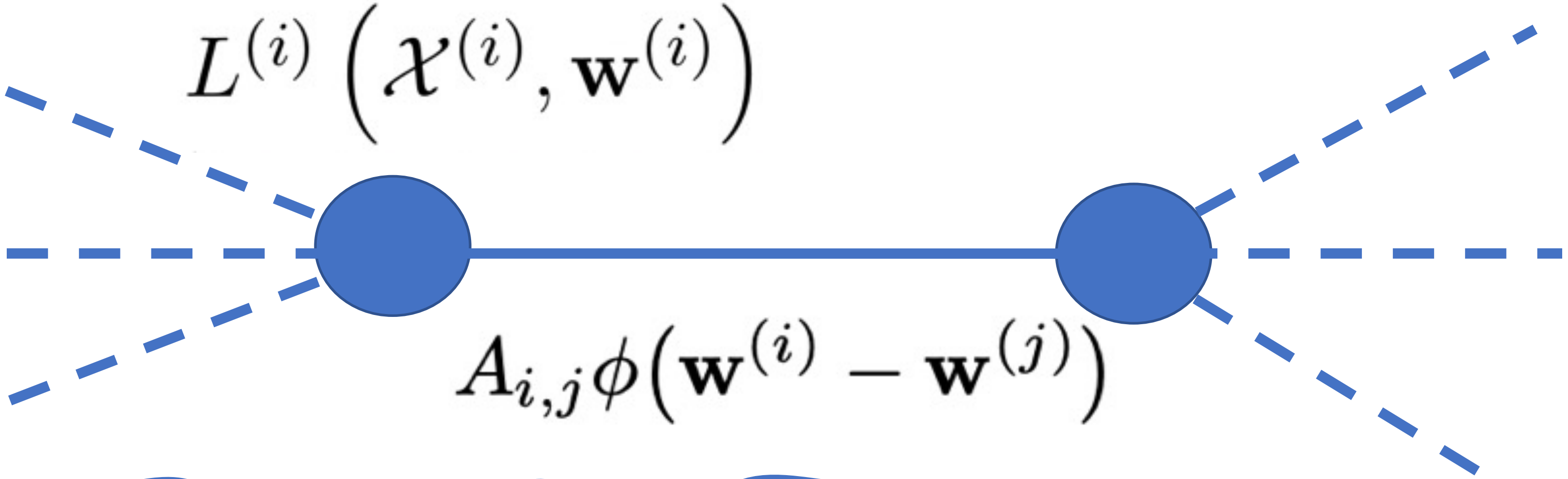
node-wise
primal update:

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

edge-wise
dual update:

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

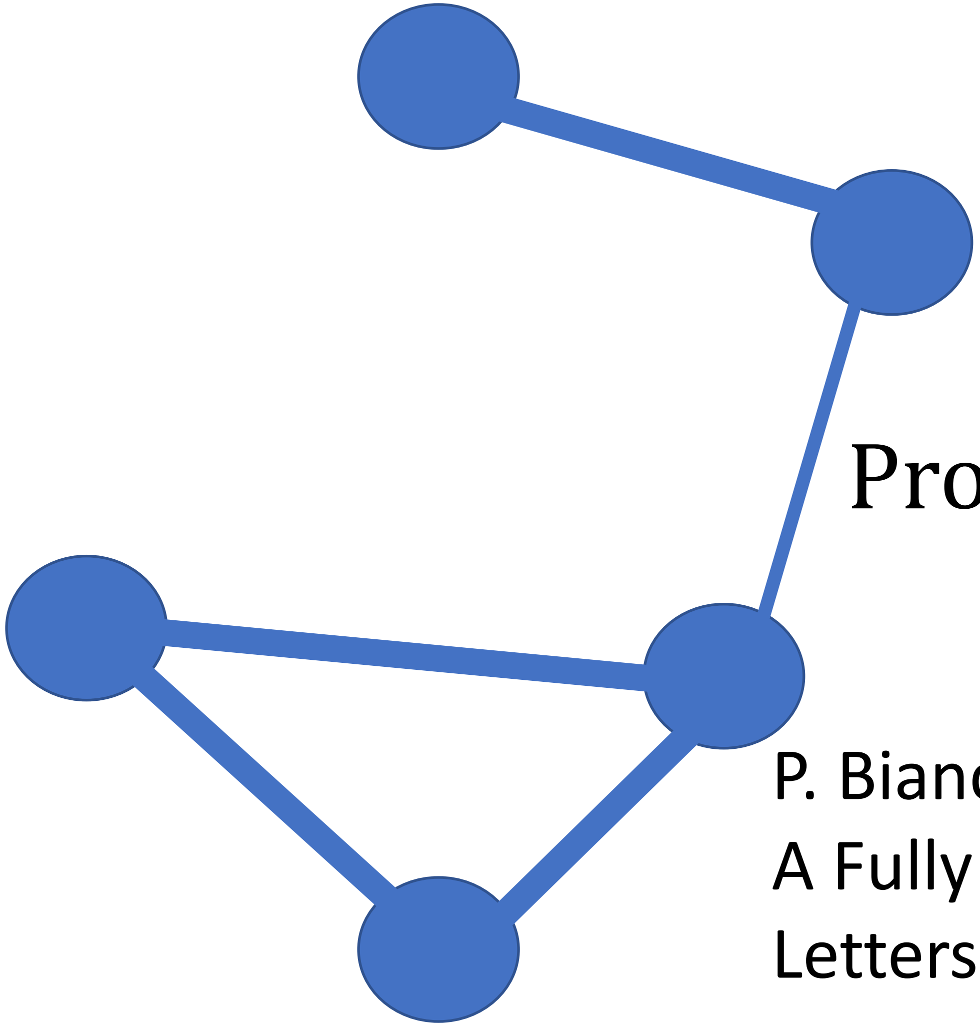
Spreading Local Results.



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

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

Networked Data as RVs

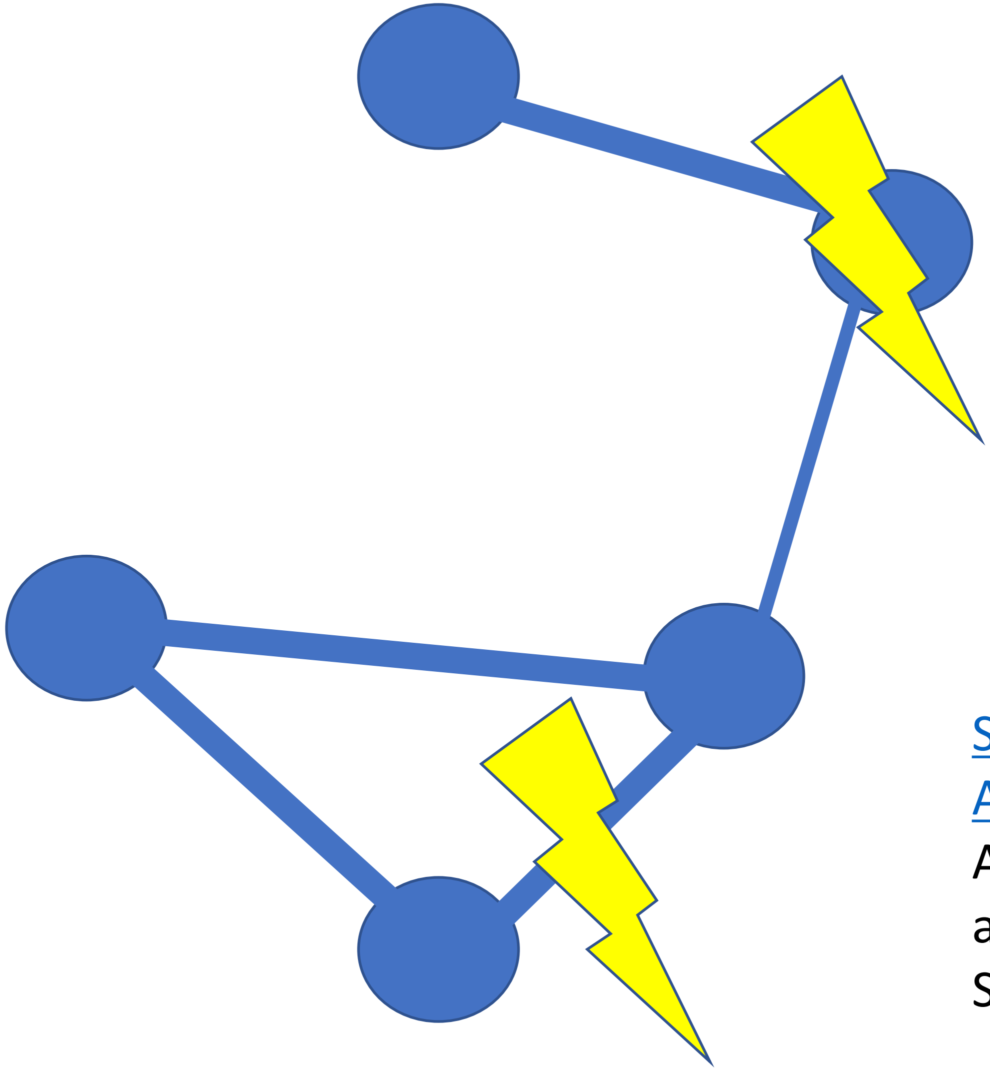


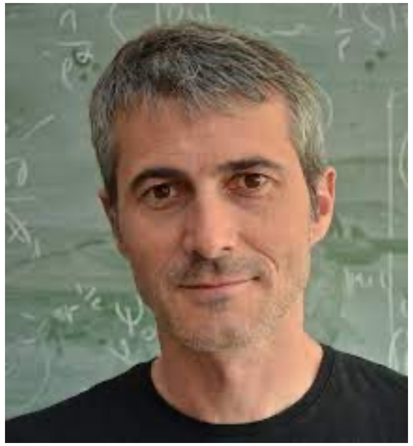
$$\mathcal{X}^{(i)} \text{ iid } \sim p^{(i)}(\mathbf{z})$$

$$\mathrm{Prob}\left(A_{i,j} = 1\right) = p$$

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

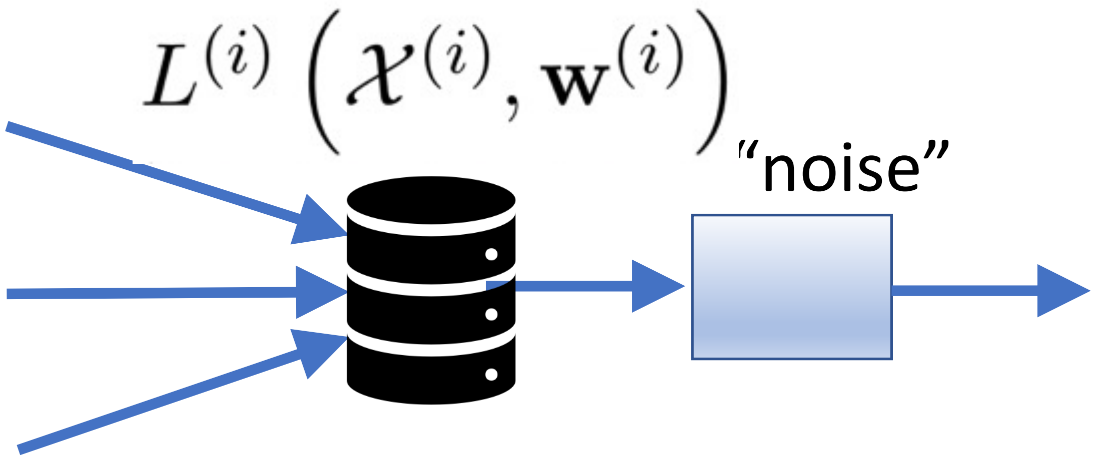
Random Node/Link Failures.

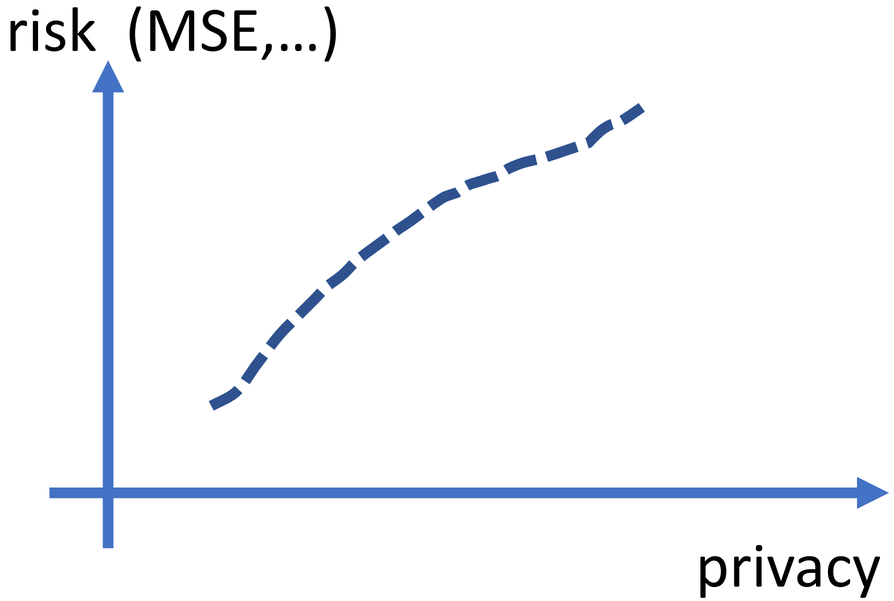




[Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications]
Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb
SIAM Journal on Optimization 2018 28:4, 2783-2808

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.

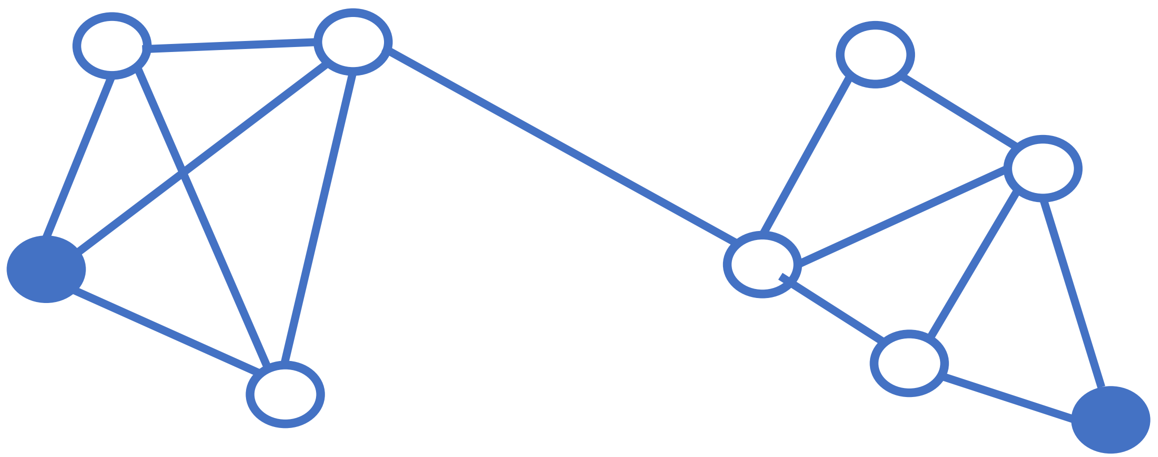
Bottom Line.

PD method solves GTVMin in distributed, robust and privacy-friendly way

….., however ….

Are GTVMin Solutions Any Good?

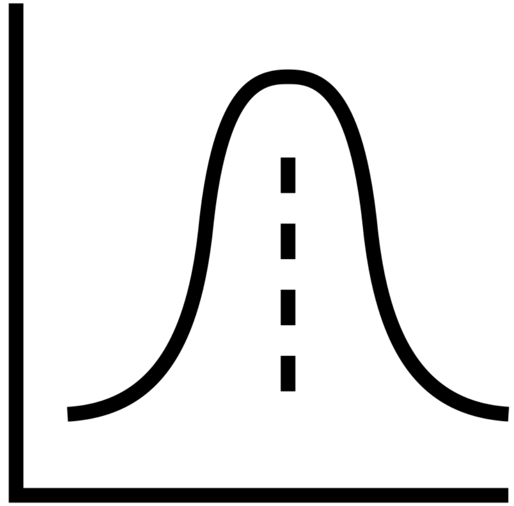
$$\min_{\mathbf{w}} \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$



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 of GTVMin



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)

Statistical Aspects.

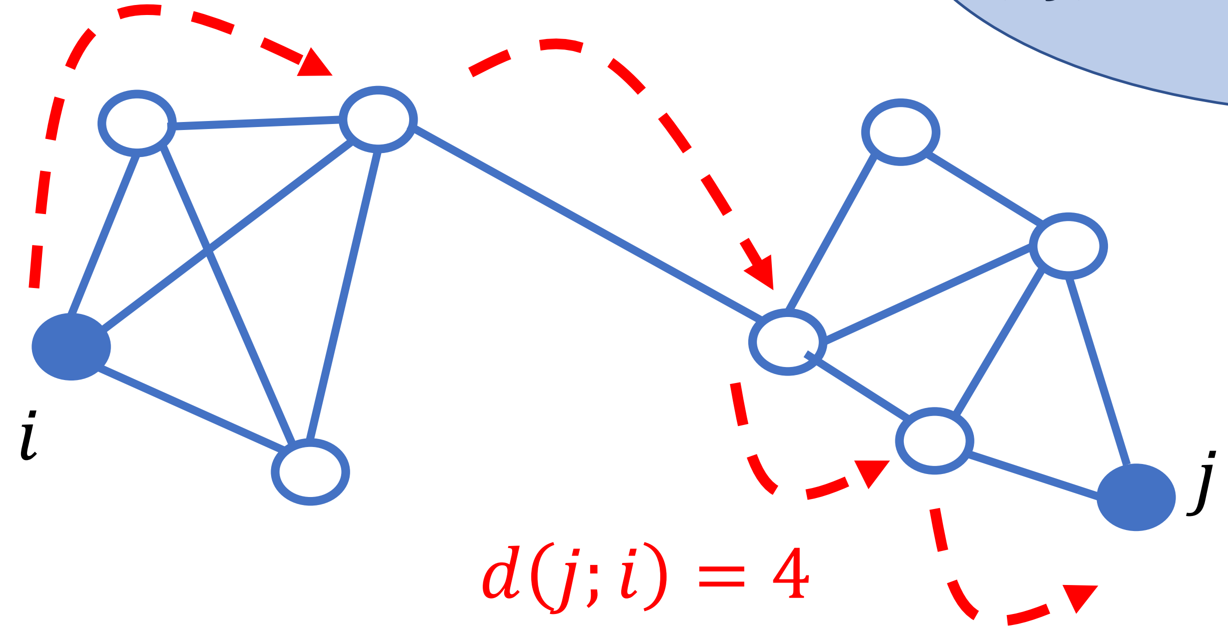
$$min_w \sum_{i \epsilon \mathrm{M}} L^{(i)}\left(w^{(i)}\right) + \lambda \sum_{\{i,j\}} A_{i,j} \phi\left(w^{(i)} - w^{(j)}\right)$$

statistical properties of GTVMin solutions?

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

Signal Processing Perspective.

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

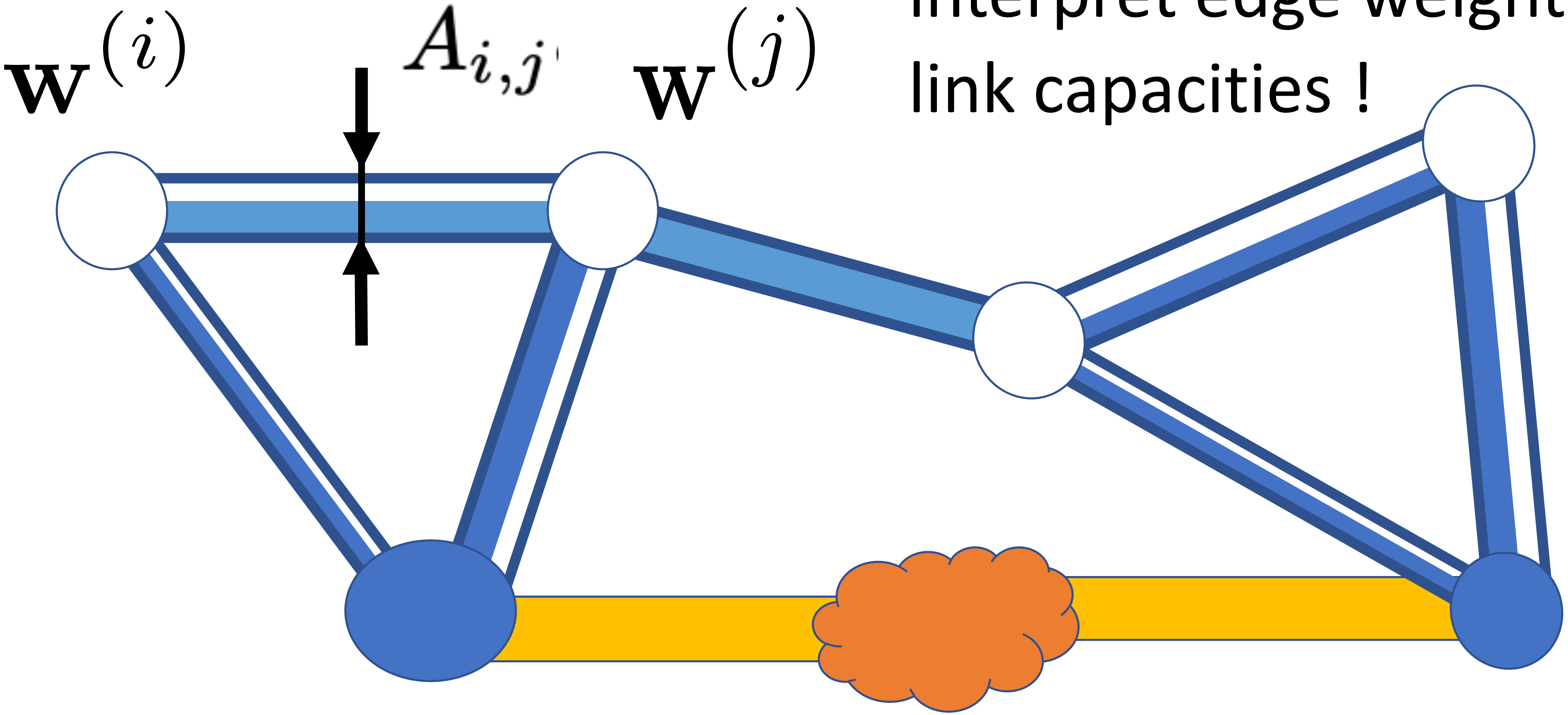


constraints
effective “bandwidth”
of graph signal $w^{(i)}$

M. Tsitsvero, S. Barbarossa and P. Di Lorenzo, "Signals on Graphs: Uncertainty Principle and Sampling," in *IEEE Transactions on Signal Processing*, 2016,

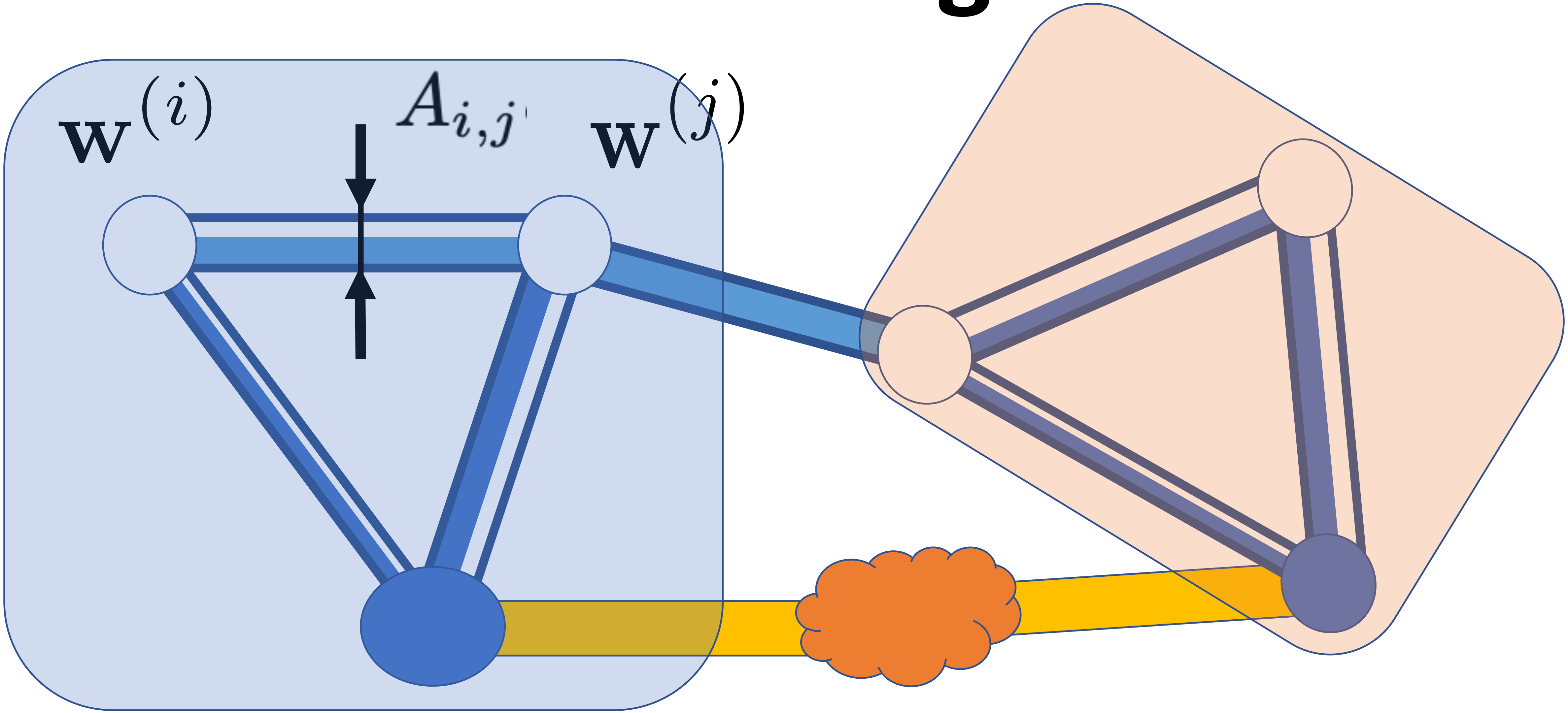
Our Perspective: Flows.

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



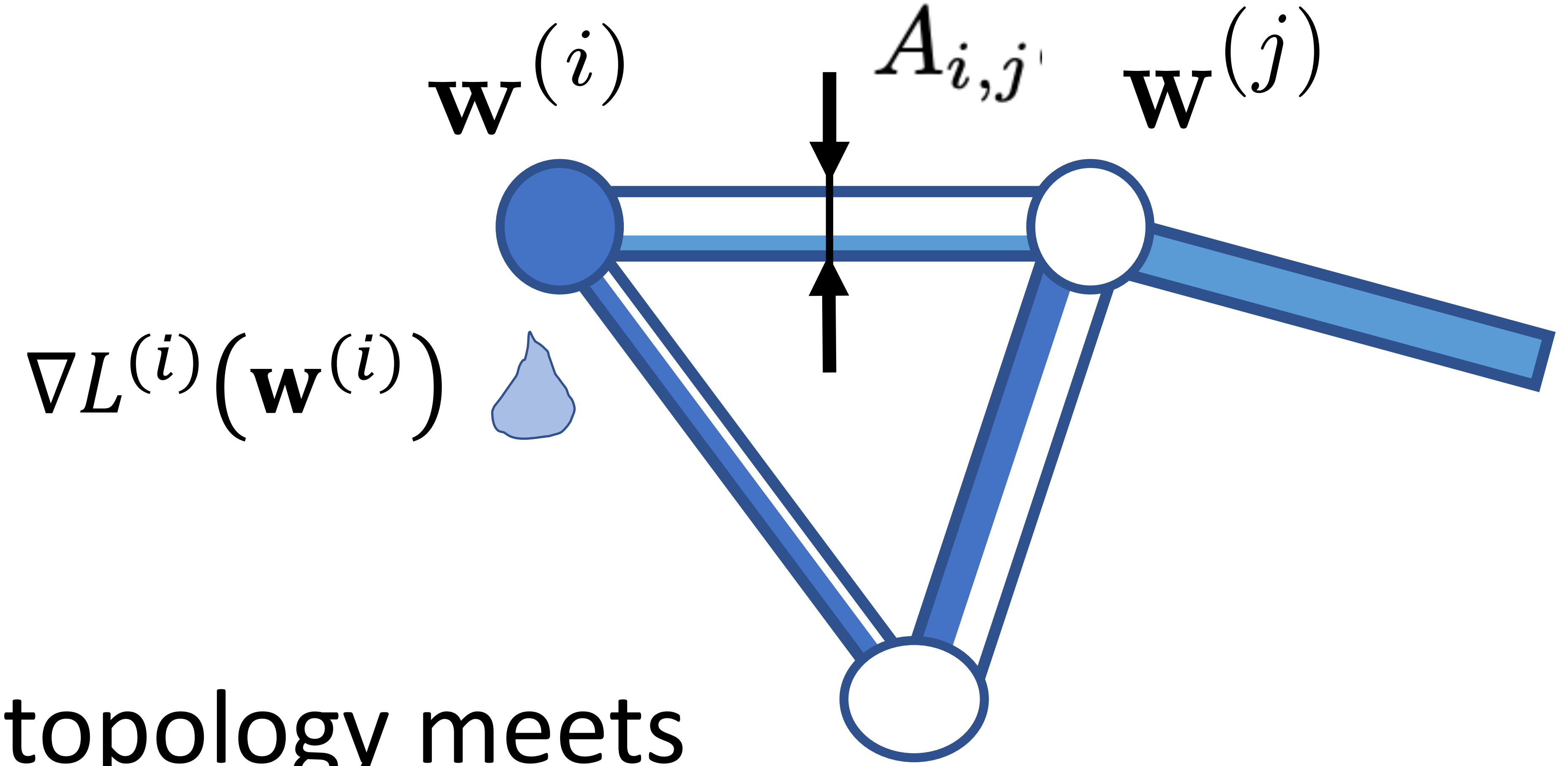
A. Jung, "On the Duality Between Network Flows and Network Lasso," in *IEEE Signal Processing Letters*, vol. 27, pp. 940-944, 2020.

Cluster-wise Pooling.



parameter vectors can only change over saturated links

Leaky Training Set.



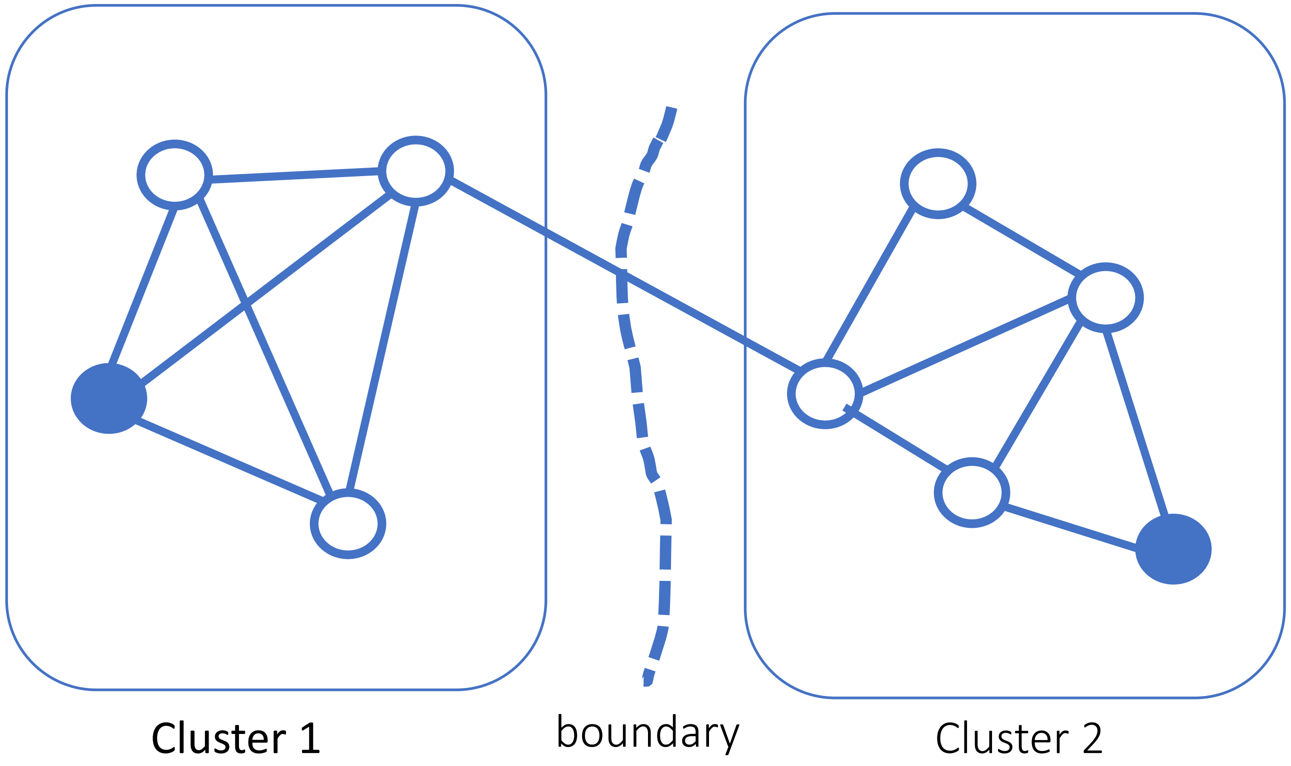
network topology meets
geometry of loss functions !

Personalization vs. Globalization

small lambda -→ pooling reduces to single local datasets
→ everybody is a single cluster

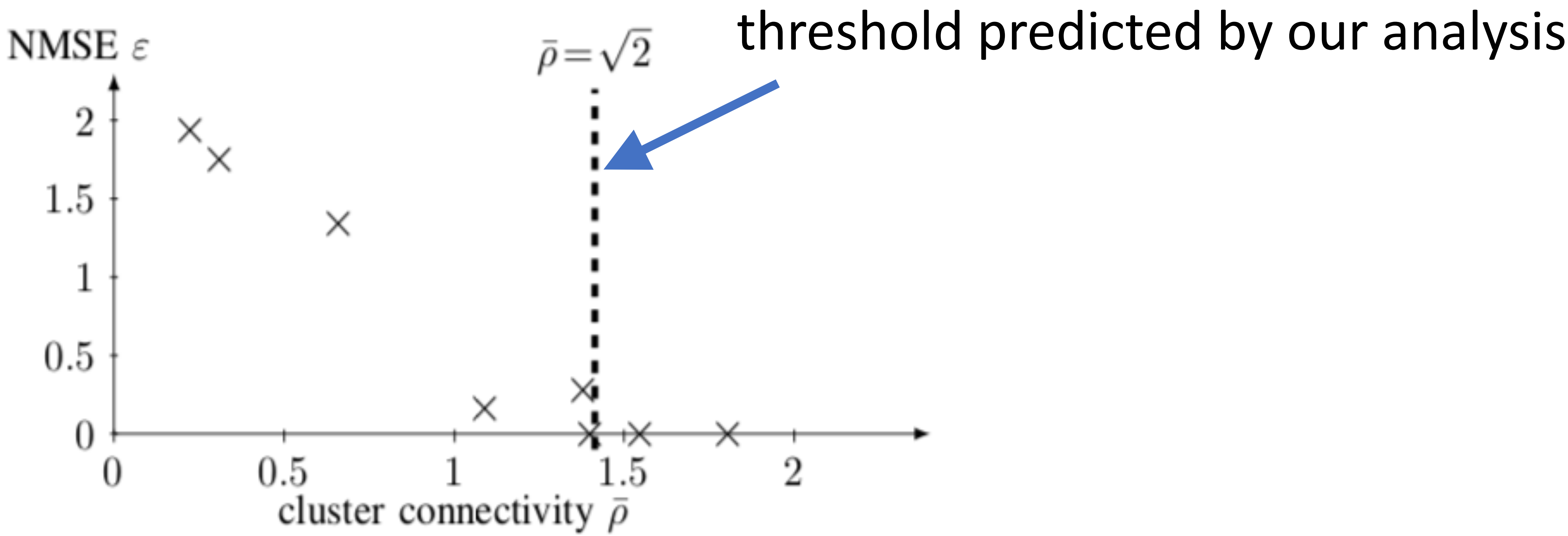
large lambda → pooling more and more local datasets
→ everybody assigned to larger and larger cluster

Measure Connectivity by Flows.



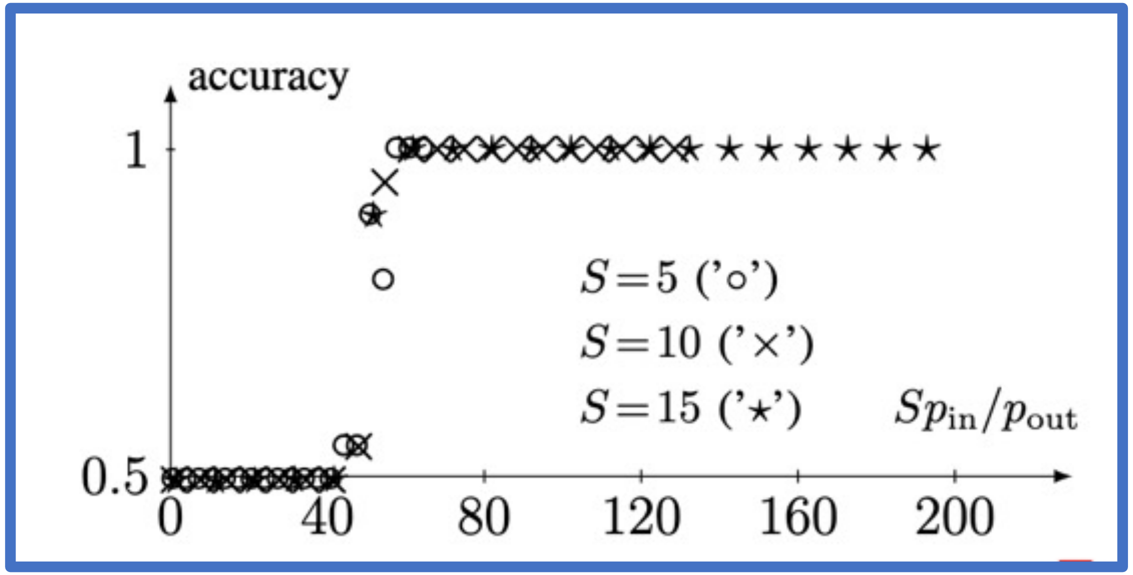
connectivity measured by flow ρ that can be routed over boundary edge

Statistical Error vs. Connectivity.



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.

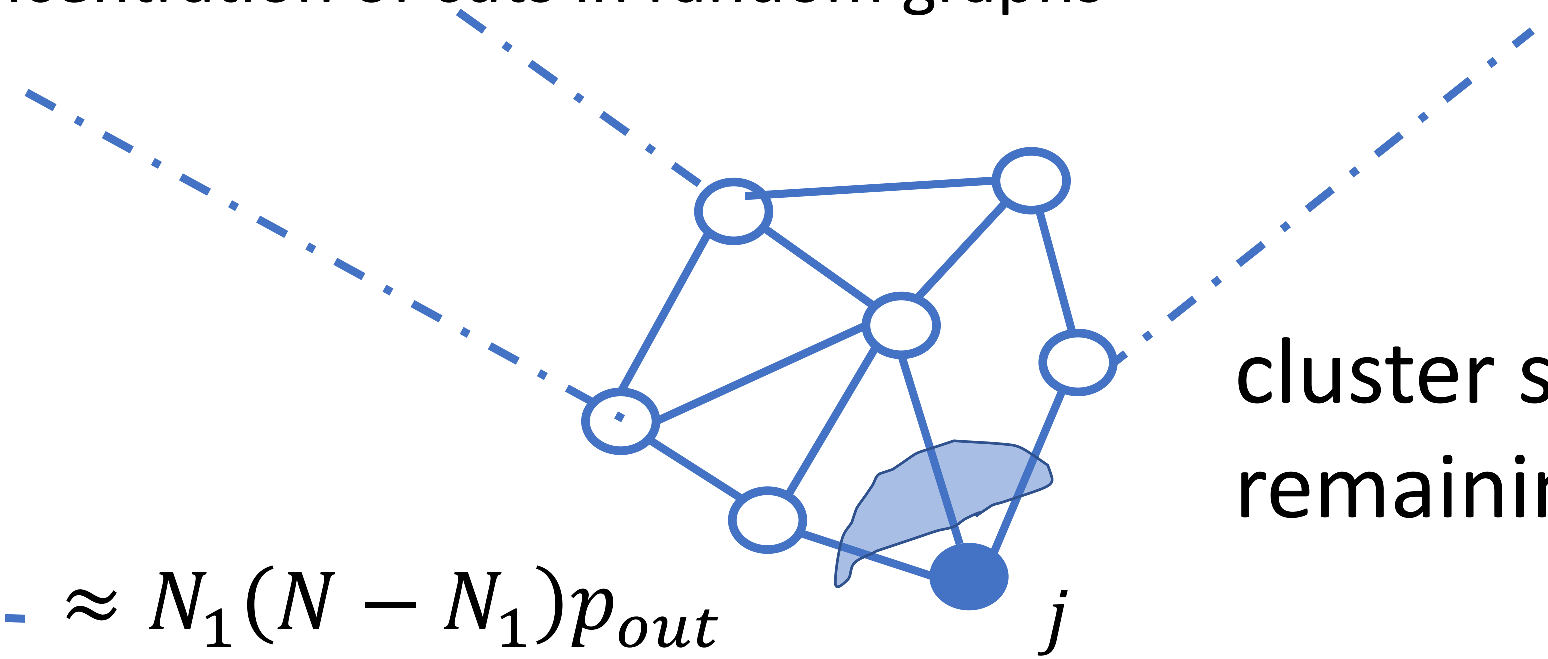


- intra-cluster edge prob p_{in}
- inter-cluster edge prob p_{out}
- S training nodes in each cluster
- critical value for S*pin/pout

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

Mathematical Device.

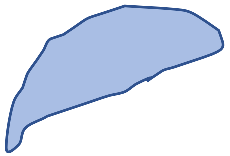
- flow conservation/Hoffman’s circulation theorem
- concentration of cuts in random graphs



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.

- GTVMin paradigm for NFL
- Dual of GTVMin is Network Flow Optim.
- solve GTVMin. with primal-dual method
- scalable and robust implementation as message passing
- GTV min. adaptively pools similar datasets

Thank you for
your attention!

Interested in doing a Phd in the happiest country in the world ?

happiest country in the world

All Images News Maps Videos ⋮ Mo

About 26.100.000 results (1,15 seconds)

Finland

1. **Finland**. For the fifth year in a row, Finland is number one when it comes to happiness. 31 Mar 2022

https://www.visitfinland.com/en/