Big Data Analytics for Autonomous Energy Grids

Georgios B. Giannakis

Acknowledgments: Profs. V. Kekatos (VaTech), S. J. Kim (UMBC), G. Mateos (UR)
H. Zhu (UT Austin); and D. Berberidis, P. Traganitis, G. Wang
NSF 1423316, 1442686, 1508993, 1509040, 1514056, 1711471

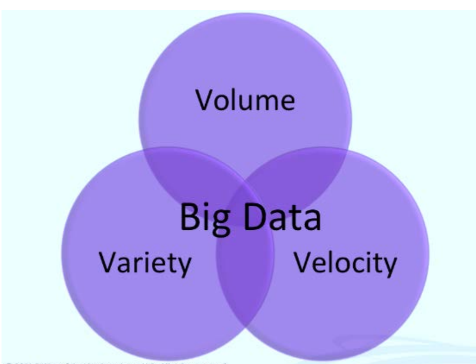
University of Minnesota
Driven to Discover[SM]

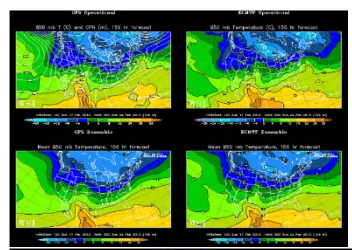
September 13, 2017

Learning from “Big Data”

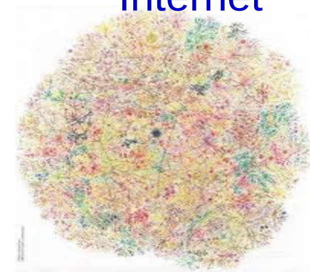
- Challenges
  - Big size ($D \gg$ and/or $N \gg$)
  - Fast streaming
  - Incomplete
  - Noise and outliers

- Opportunities in key tasks
  - Dimensionality reduction
  - Online and robust regression, classification and clustering
  - Denoising and imputation

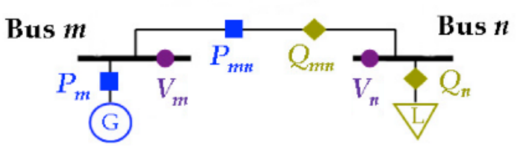




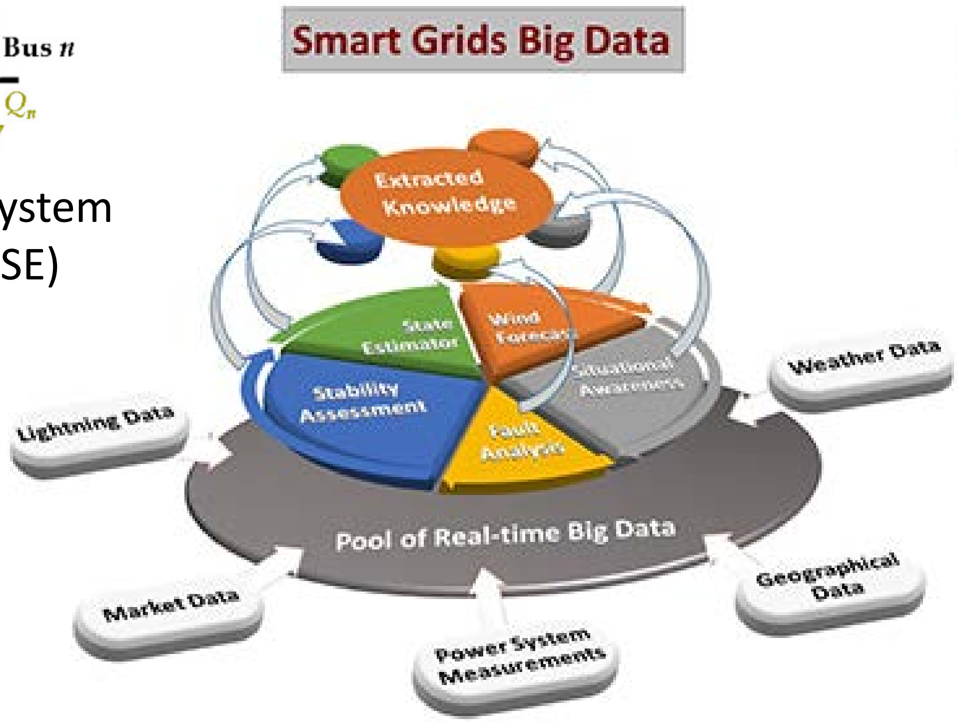
Internet



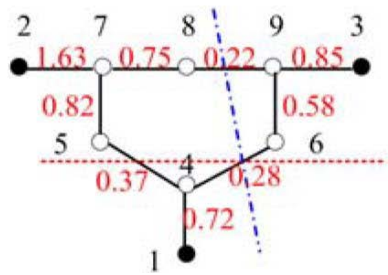
Analytics for energy grids



Distributed power system state estimation (PSSE)



Bad data cleansing



Controlled islanding

Load and price forecasting

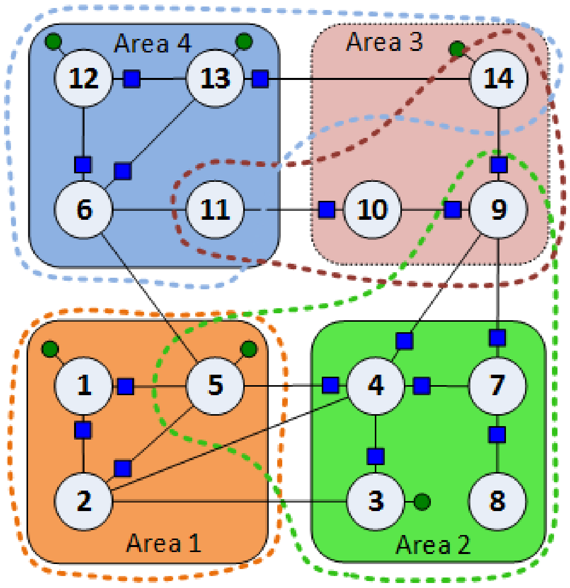
Consumer profiling

Scalable clustering into cellular blocks for **autonomous energy grids** (AEGs)

Roadmap

- Context and motivation
- Estimation with big data
  - Distributed, robust, and scalable PSSE
  - Sketching, censoring, and tracking
  - Spatio-temporal imputation and forecasting
- Large-scale data and graph clustering
- Closing comments

Centralized PSSE



- AEG with K cells $\mathcal{N} = \bigcup_{k=1}^{K} \mathcal{N}_k$

  bus voltages $\mathbf{v} := [V_1, \ldots, V_N]^T \in \mathbb{C}^N$

- SCADA measurements (quadratic in **v**)

$$z_k^\ell = h_k^\ell(\mathbf{v}) + \epsilon_k^\ell, \quad \forall k, \ell$$

- Nonlinear least-squares (LS) state estimator

$$\hat{\mathbf{v}} := \arg\min_{\mathbf{v}} \sum_{k=1}^{K} \sum_{\ell=1}^{M_k} [z_k^\ell - h_k^\ell(\mathbf{v})]^2 \qquad \textbf{(C-SE)}$$

- Gauss-Newton iterative solvers via linearization, e.g., [Abur-Exposito'04] sensitive to initialization (esp. w/ fast-varying states): convergence?

Convexification via SDR

- **Trick**: make z_k^ℓ linear in $\mathbf{V} := \mathbf{v}\mathbf{v}^{\mathcal{H}}$

$$z_k^\ell = h_k^\ell(\mathbf{v}) + \epsilon_k^\ell = \mathrm{Tr}(\mathbf{H}_k^\ell \mathbf{V}) + \epsilon_k^\ell$$

$$\hat{\mathbf{V}} := \arg\min_{\mathbf{V}} \sum_{k=1}^{K} \sum_{\ell=1}^{M_k} \left[z_k^\ell - \mathrm{Tr}(\mathbf{H}_k^\ell \mathbf{V})\right]^2$$

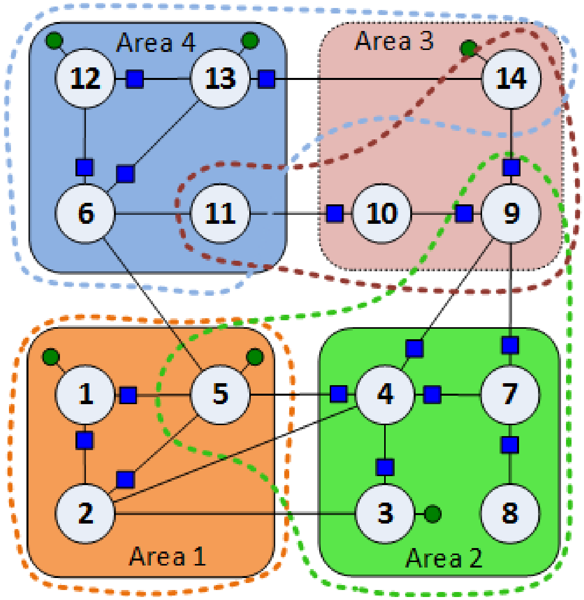
s.to $\mathbf{V} \succeq \mathbf{0}$, and $\cancel{\mathrm{rank}(\mathbf{V}) = 1}$ **(C-SDP)**

- SDR for SE [Zhu-GG'11] for SE; SDR for OPF [Bai etal'08], [Lavaei-Low'11]
  - Generalizations include PMU data, and robust SDR-based state estimation
  - (Near-)optimal regardless of initialization; polynomial complexity $O(N^{4.5}\log(1/\epsilon))$

Desiderata: Decentralized SDR scalable with control area size, privacy-preserving, and solvable at affordable communication cost

H. Zhu and G. B. Giannakis, ``Estimating the state of AC power systems using semi-definite programming," in *Proc. North American Power Symposium*, Boston, MA, Aug. 2011.

Cost decomposition



- Include tie-line buses; split local LS cost per $\mathcal{N}_{(k)}$

$$f_k(\mathbf{V}_{(k)}) := \sum_{\ell=1}^{M_k} \left[z_k^\ell - \mathrm{Tr}(\mathbf{H}_{(k)}^\ell \mathbf{V}_{(k)}) \right]^2$$

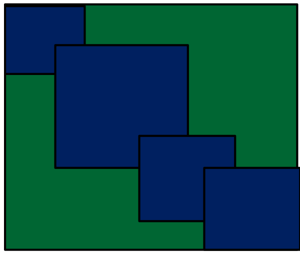
$$\mathcal{N}_{(2)} := \mathcal{N}_2 \cup \{5, 9\}$$

$$\hat{\mathbf{V}} := \arg\min_{\mathbf{V}} \sum_{k=1}^{K} \sum_{\ell=1}^{M_k} \left[z_k^\ell - \mathrm{Tr}(\mathbf{H}_k^\ell \mathbf{V}) \right]^2 \quad \text{s.to} \quad \mathbf{V} \succeq \mathbf{0}$$

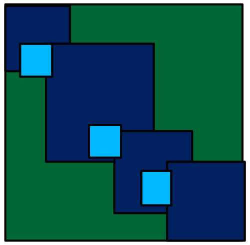
$$\hat{\mathbf{V}} := \arg\min_{\mathbf{V}} \sum_{k=1}^{K} f_k(\mathbf{V}_{(k)}) \quad \text{s.to} \quad \mathbf{V} \succeq \mathbf{0}$$

Challenge: as $\{\mathcal{N}_{(k)}\}$ overlap partially, PSD const. couples $\{\mathbf{V}_{(k)}\}$

Blessing: overlap $\longrightarrow$ global; no overlap: $\mathbf{V} \succeq \mathbf{0} \Leftrightarrow \mathbf{V}_{(k)} \succeq \mathbf{0}, \quad \forall k$



Decentralized SDR for PSSE



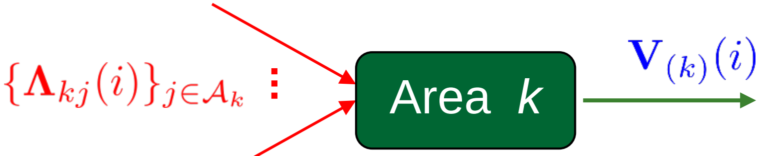
- If graph (w/ areas as nodes, overlaps as edges) is a tree, then

$$\hat{\mathbf{V}} := \arg\min_{\mathbf{V}} \sum_{k=1}^{K} f_k(\mathbf{V}_{(k)}) \quad \text{s.to} \quad \mathbf{V} \succeq \mathbf{0} \qquad \textbf{(C-SDP)}$$

$$\Longleftrightarrow$$

$$\{\hat{\mathbf{V}}_{(k)}\} := \arg\min_{\{\mathbf{V}_{(k)}\}} \sum_{k=1}^{K} f_k(\mathbf{V}_{(k)}) \quad \text{s.to} \quad \mathbf{V}_{(k)} \succeq \mathbf{0}, \quad \mathbf{V}_{(k)}^{[j]} = \mathbf{V}_{(j)}^{[k]}$$

- ADMM [Glowinski-Marrocco'75]; for D-Estimation [Schizas-Giannakis'06]
  - Iterates between local variables and multipliers per equality constraint

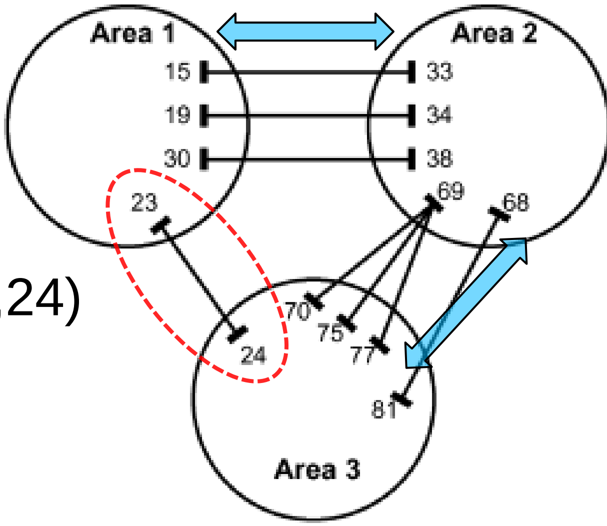


Local SDP

Linear Update

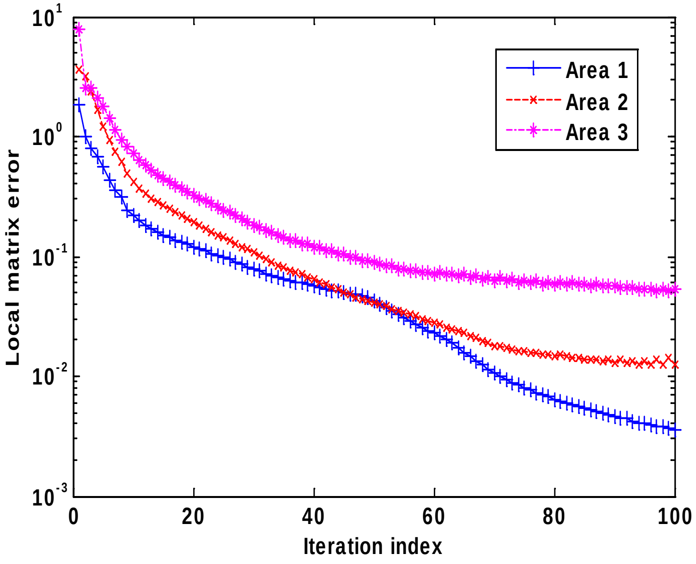
- Converges $\mathbf{V}_{(k)}(i) \rightarrow \hat{\mathbf{V}}_{(k)}$ even for noisy-async. links [Schizas-GG'08], [Zhu-GG'09]

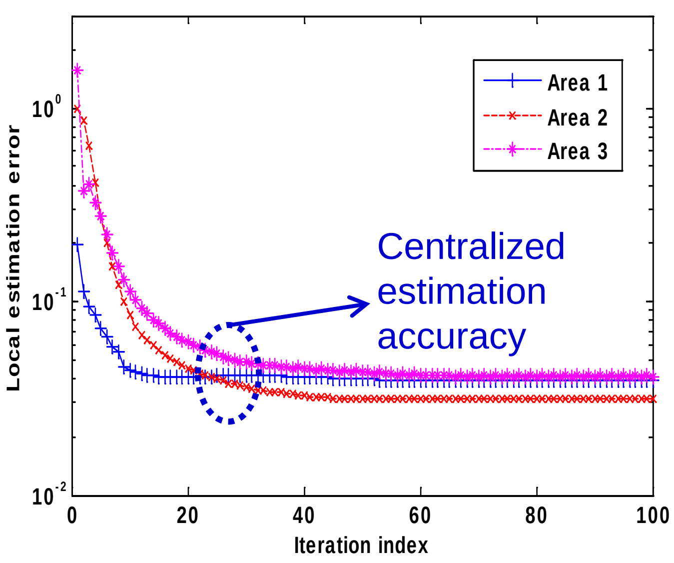
118-bus test case



- Triangular configuration [Min-Abur'06]
- Power flow meters on all tie lines except for (23,24)

➡ graph of areas is a tree





- Local norms

$$\|\mathbf{v}_{(k)}(i) - \mathbf{v}_{(k)}\|_2$$

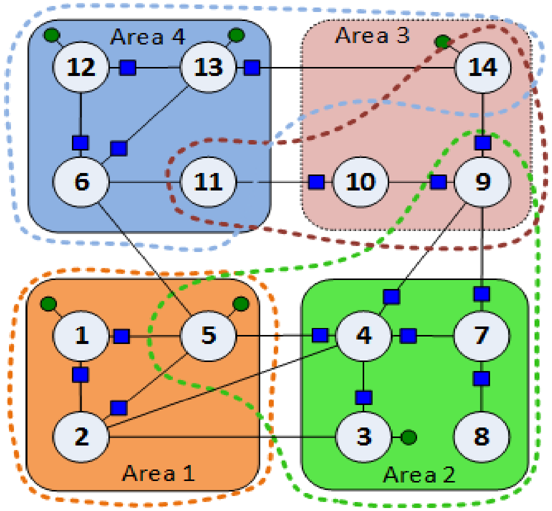
converge in only
20 iterations!

H. Zhu and G. B. Giannakis, ``Power system nonlinear state estimation using distributed semi-definite programming," *IEEE Journal on Special Topics in Signal Processing*, pp. 1039-1050, December 2014.

Decentralized PSSE for linear models

- Local linear(ized) model $\mathbf{z}_k = \mathbf{H}_{(k)}\mathbf{v}_{(k)} + \mathbf{n}_k$
- Regional PSSEs $\min_{\mathbf{v}_{(k)} \in \mathcal{V}_k} f_k(\mathbf{v}_{(k)})$
- Coupled local problems

$$\min_{\{\mathbf{v}_{(k)}\}} \sum_{k=1}^{K} f_k(\mathbf{v}_{(k)})$$
$$\text{s.to} \quad \mathbf{v}_{(k)}[l] = \mathbf{v}_{(l)}[k]$$



$\mathcal{S}_2 := \mathcal{N}_{(2)} \setminus \mathcal{N}_2$

S1. $\mathbf{v}_{(k)}^{t+1} = \arg \min_{\mathbf{v}_{(k)} \in \mathcal{V}_k} f_k(\mathbf{v}_{(k)}) + \frac{c}{2} \sum_{i \in \mathcal{S}_k} \left(v_{(k)}(i) - \mu_k^t(i) \right)^2$

S2. $\mu_k^{t+1}(i) = \mu_k^t(i) + \left(v_{(l)}^{t+1}[i] - \frac{v_{(k)}^t(i) + v_{(l)}^t[i]}{2} \right)$

- ADMM solver: convergent w/ minimal $(\text{no } \boldsymbol{\mu}_k)$ exchanges and privacy-preserving

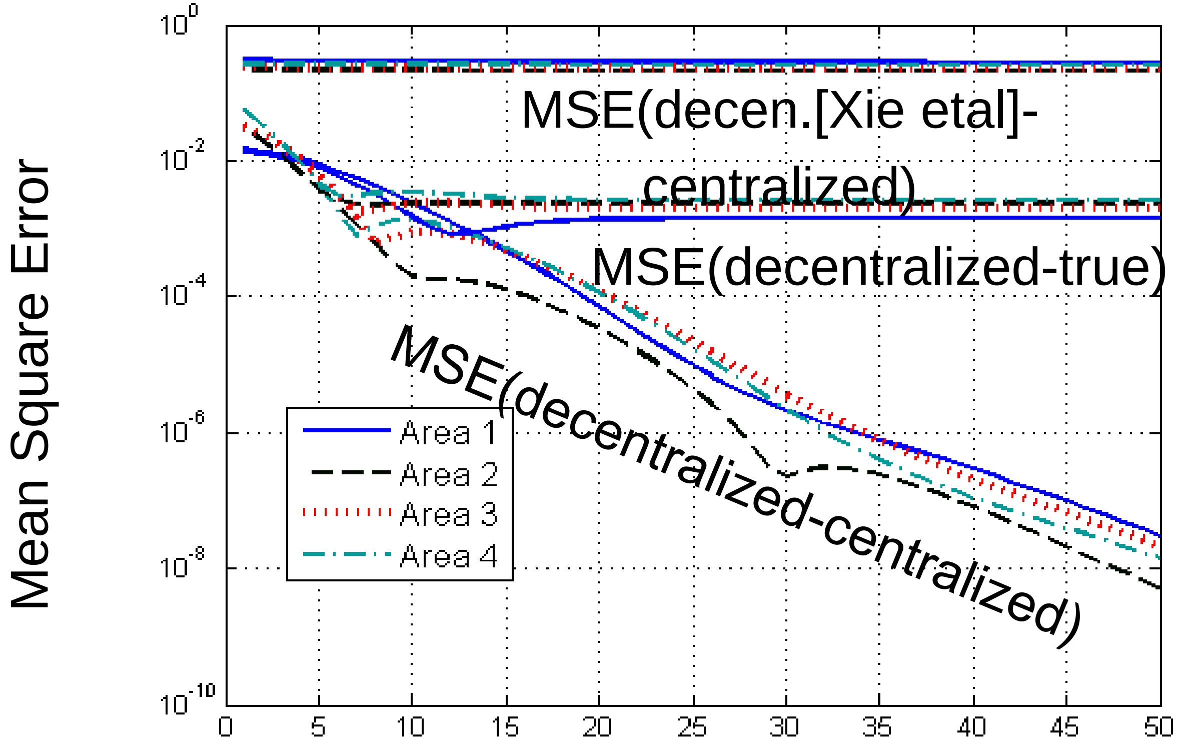
V. Kekatos and G. B. Giannakis, "Distributed Robust Power System State Estimation," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1617-1626, May 2013.

Simulated test

S1. $\mathbf{v}_{(k)}^{t+1} = \left(\mathbf{H}_{(k)}^T \mathbf{H}_{(k)} + c \cdot \mathbf{D}_k\right)^{-1} \left(\mathbf{H}_{(k)}^T \mathbf{z}_k + c \cdot \mathbf{D}_k \boldsymbol{\mu}_k^t\right), \quad [\mathbf{D}_k]_{ii} = |\mathcal{S}_k^i|$

S2 $\mu_k^{t+1}(i) = \mu_k^t(i) + \left(v_{(l)}^{t+1}[i] - \frac{v_{(k)}^t(i) + v_{(l)}^t[i]}{2}\right)$

.



L. Xie, C. Choi, and S. Kar, ``Cooperative distributed state estimation: Local observability relaxed,'' in *Proc. IEEE PES General Meeting*, Detroit, MI, July 2011.

Decentralized bad data cleansing

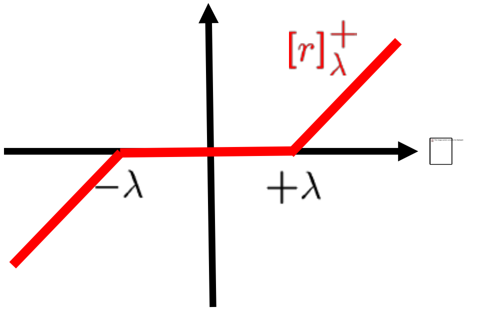
$$\mathbf{z} = \mathbf{H}\mathbf{v} + \mathbf{n} + \mathbf{o}$$

- Reveal *single* and *block* outliers via

$$f(\mathbf{v}) := \min_{\mathbf{o}} \frac{1}{2}\|\mathbf{z} - \mathbf{H}\mathbf{v} - \mathbf{o}\|_2^2 + \lambda\|\mathbf{o}\|_1$$
$$= \sum_{m=1}^{M} h(z_m - \mathbf{h}_m^T \mathbf{v})$$

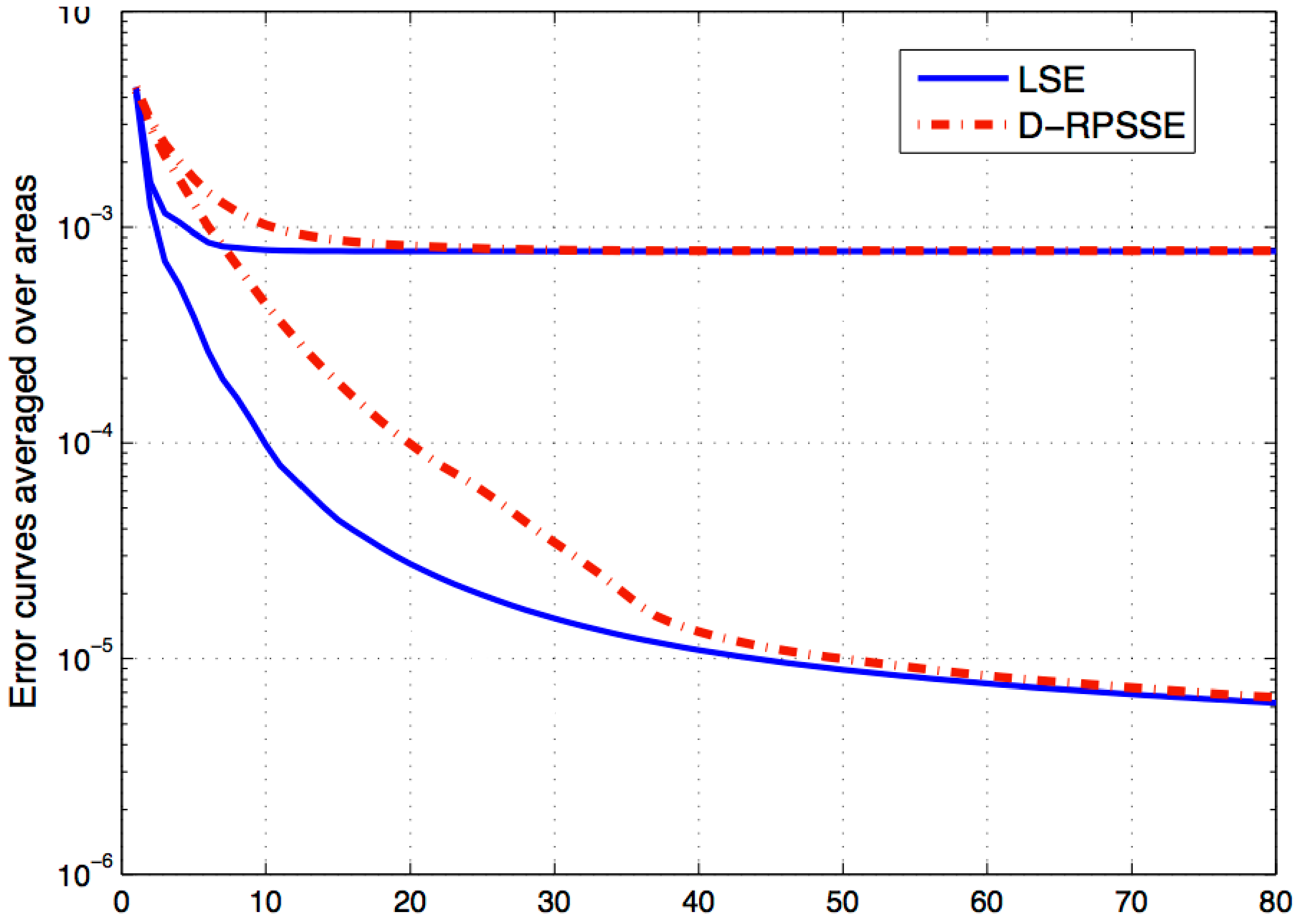
S1. $\mathbf{v}_{(k)}^{t+1} = \left(\mathbf{H}_{(k)}^T \mathbf{H}_{(k)} + c \cdot \mathbf{D}_k\right)^{-1} \left(\mathbf{H}_{(k)}^T (\mathbf{z}_k - \mathbf{o}_k) + c \cdot \mathbf{D}_k \boldsymbol{\mu}_k^t\right)$

S2. $\mathbf{o}_k^{t+1} = \left[\mathbf{z}_k - \mathbf{H}_{(k)} \mathbf{v}_{(k)}^{t+1}\right]_\lambda^+$



S3. $\mu_k^{t+1}(i) = \mu_k^t(i) + \left(v_{(l)}^{t+1}[i] - \frac{v_{(k)}^t(i) + v_{(l)}^t[i]}{2}\right)$

D-PSSE on a 4,200-bus grid



Robust LAV

- Robustness to outliers via least-absolute-value (LAV) criterion

$$\boxed{\min_{\boldsymbol{v}\in\mathbb{C}^n} \sum_{\ell=1}^{L} |z_\ell - h_\ell(\boldsymbol{v})|}$$

nonconvex, non-smooth!

2017 Ukrainian blackout
by cyberattacks

- Existing approaches (slow and non-scalable)
  - Subgradient solver [Jabr-Pal'03]
  - Successive linear programming [Abur and Celik'91]

- Deterministic solver via composite optimization [Wang-Giannakis-Chen'17]

$$\boxed{\boldsymbol{v}^{t+1} := \arg\min_{\boldsymbol{v}} \left\{ \sum_\ell \left| \boldsymbol{h}_\ell^{\mathcal{H}} \boldsymbol{v} + \tilde{z}_\ell \right| + \frac{1}{2\mu_t} \left\| \boldsymbol{v} - \boldsymbol{v}^t \right\|_2^2 \right\}}$$

Locally tight quadratic upper bound; convex!

Linearization of $h_\ell(\boldsymbol{v})$ around iterate $\boldsymbol{v}^t$

Constant depending on z_ℓ and $h_\ell(\boldsymbol{v}^t)$

S. Lewis and S. J. Wright, "Proximal method for composite minimization," *Mathematical Programming*, vol. 158, no. 1-2, pp. 501–546, July 2016.

Scalable stochastic solver

- Stochastic composite optimization

$$\boldsymbol{v}^{t+1} := \arg\min_{\boldsymbol{v}} \left\{ \left| \boldsymbol{h}_{\ell_t}^{\mathcal{H}} \boldsymbol{v} + \tilde{z}_{\ell_t} \right| + \frac{1}{2\mu_t} \left\| \boldsymbol{v} - \boldsymbol{v}^t \right\|_2^2 \right\}$$

Draw datum $\ell_t \in \{1, \ldots, L\}$ randomly per *t*

Process one datum per *t*

- Closed-form updates

$$\boldsymbol{v}^{t+1} = \boldsymbol{v}^t + \mathrm{Proj}_{\mu_t}(\tilde{z}'_{\ell_t} / \|\boldsymbol{h}_{\ell_t}\|_2^2) \cdot \boldsymbol{h}_{\ell_t}$$

Projection onto interval $[-\mu_t, \mu_t]$

- Merits
  - Very few operations per iteration (due to highly sparse $\boldsymbol{h}_\ell$ vectors)
  - Fast linear convergence under suitable conditions [Duchi-Feng'17]
  - Further acceleration via mini-batching of non-overlapping measurements

- ~5 mins on desktop for 9,241-bus grid; not enough memory for Gauss-Newton

G. Wang, G. B. Giannakis, and J. Chen, "Robust and Scalable Power System State Estimation via Composite Optimization," *arXiv:1708.06013*, 2017.

Roadmap

- Context and motivation
- Estimation with big data
  - Distributed, robust, and scalable PSSE
  - Sketching, censoring, and dynamic tracking
  - Spatio-temporal imputation and forecasting
- Large-scale data and graph clustering
- Closing comments

Random projections for data sketching

Least-squares (LS) with PMU data Given $\mathbf{y} \in \mathbb{R}^D, \; \mathbf{X} \in \mathbb{R}^{D \times p}$

$$\boldsymbol{\theta}_{\mathrm{LS}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

If $\mathrm{rank}(\mathbf{X}) = p \Rightarrow \boldsymbol{\theta}_{\mathrm{LS}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$

- SVD incurs complexity $\mathcal{O}(Dp^2)$ **Q:** What if $D \gg p$?
- LS estimate via (pre-conditioning) **random projection** matrix $\mathbf{R}_{d \times D}$

$$\check{\boldsymbol{\theta}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\overbrace{\mathbf{S}_d \mathbf{H}_D \mathbf{B}_D}^{\mathbf{R}} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_2^2 \qquad d \ll D$$

- For $d = \mathcal{O}(p \log p \cdot \log D + \epsilon^{-1} D \log p)$, complexity reduces to $o(Dp^2)$

M. W. Mahoney, Randomized Algorithms for Matrices and Data. *Foundations and Trends In Machine Learning*, vol. 3, no. 2, pp. 123-224, Nov. 2011.

Performance of randomized LS

- Based on the Johnson-Lindenstrauss lemma [JL'84]

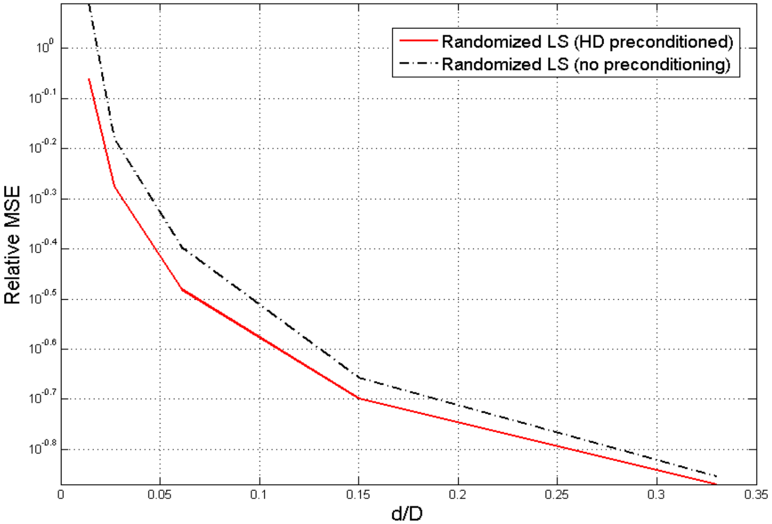
Theorem. For any $\epsilon > 0$, if $d = \mathcal{O}(p \log p / \epsilon^2)$ then w.h.p.

$$\|\mathbf{y} - \mathbf{X}\check{\boldsymbol{\theta}}_{\text{LS}}\|_2 \leq (1+\epsilon)\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_{\text{LS}}\|_2$$

$$\|\boldsymbol{\theta}_{\text{LS}} - \check{\boldsymbol{\theta}}_{\text{LS}}\|_2 \leq \sqrt{\epsilon}\,\kappa(\mathbf{X})\sqrt{\gamma^{-2} - 1}\,\|\boldsymbol{\theta}_{\text{LS}}\|_2$$

$\kappa(\mathbf{X})$ condition number of $\mathbf{X}$; and $\gamma = \|\hat{\mathbf{y}}\|_2 / \|\mathbf{y}\|_2$

- Uniform sampling versus Hadamard preconditioning
  - *D* = 10,000 and *p* =50
  - Performance depends on ***X*** and ***y***



D. P. Woodruff, ``Sketching as a Tool for Numerical Linear Algebra,'' *Foundations and Trends in Theoretical Computer Science*, vol. 10, pp. 1-157, 2014.

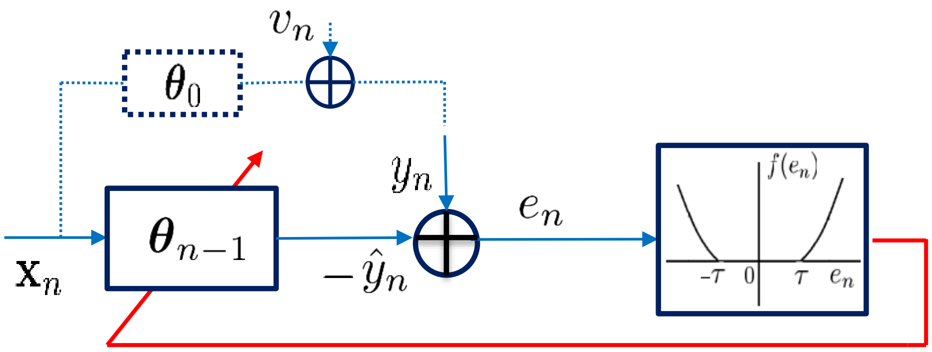
Online censoring for large-scale regressions

- **Key idea**: Sequentially test/update LS estimates **only** for informative data

- Adaptive censoring (AC) rule:

  Censor if

$$|\underbrace{y_n - \mathbf{x}_n^T \boldsymbol{\theta}_{n-1}}_{\hat{y}_n}| < \tau\sigma$$



- Criterion

$$f_n(\boldsymbol{\theta}) = f(e_n) := \begin{cases} \frac{e_n^2}{2} - \frac{\tau^2\sigma^2}{2} & |e_n| > \tau\sigma \\ 0 & |e_n| \leq \tau\sigma \end{cases}$$

- Threshold controls avg. data reduction: $\tau \approx Q^{-1}(\frac{1}{2}(1-\frac{d}{D})), \quad D \gg p$

D. K. Berberidis, G. Wang, G. B. Giannakis, and V. Kekatos, ``Adaptive Estimation from Big Data via Censored Stochastic Approximation," *Proc. of Asilomar Conf.*, Pacific Grove, CA, Nov. 2014.

Censoring algorithms and performance

- AC least mean-squares (LMS)

$$\boxed{\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + \mu(1 - c_n)\mathbf{x}_n(y_n - \mathbf{x}_n^T \hat{\boldsymbol{\theta}}_{n-1})} \qquad c_n = \begin{cases} 1, & \frac{|y_n - \mathbf{x}_n^T \boldsymbol{\theta}_{n-1}|}{\sigma} \leq \tau \\ 0, & \text{otherwise.} \end{cases}$$

- AC recursive least-squares (RLS) at complexity $\mathcal{O}(dp^2)$

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + (1 - c_n)\frac{1}{n}\hat{\mathbf{C}}_n \mathbf{x}_n (y_n - \mathbf{x}_n^T \hat{\boldsymbol{\theta}}_{n-1})$$

$$\hat{\mathbf{C}}_n = \frac{n}{n-1}\left[\hat{\mathbf{C}}_{n-1} - (1 - c_n)\hat{\mathbf{C}}_{n-1}\mathbf{x}_n \mathbf{x}_n^T \hat{\mathbf{C}}_{n-1}\left(n - 1 + \mathbf{x}_n^T \hat{\mathbf{C}}_{n-1}\mathbf{x}_n\right)^{-1}\right]$$

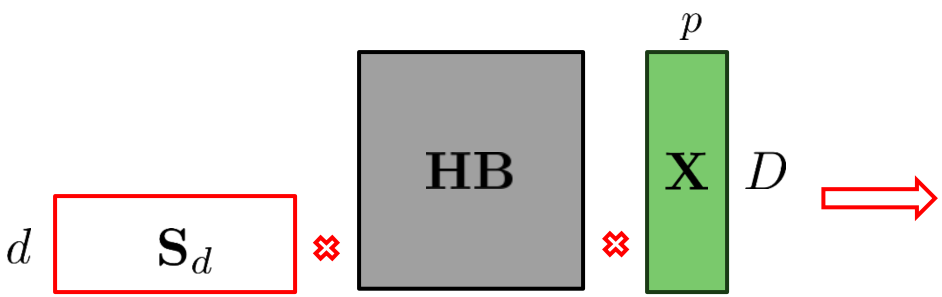
Proposition 1 **AC-RLS** $\frac{1}{n}\mathrm{tr}\left(\mathbf{R}_{\mathrm{x}}^{-1}\right)\sigma^2 \leq \mathbf{E}\left[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{1}{n}\frac{\mathrm{tr}\left(\mathbf{R}_{\mathrm{x}}^{-1}\right)\sigma^2}{2Q(\tau)} \quad \forall n \geq k$

AC-LMS $\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{\exp(4L^2/\alpha^2)}{n^2}\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2^2 + \frac{\Delta}{L^2}\right) + 8\frac{\Delta}{\alpha^2}\frac{\log n}{n}$

D. K. Berberidis, V. Kekatos, and G. B. Giannakis, ``Online Censoring for Large-Scale Regressions with Application to Streaming Big Data," *IEEE Trans. on Signal Processing*, vol. 64, pp. 3854-3867, 2016.

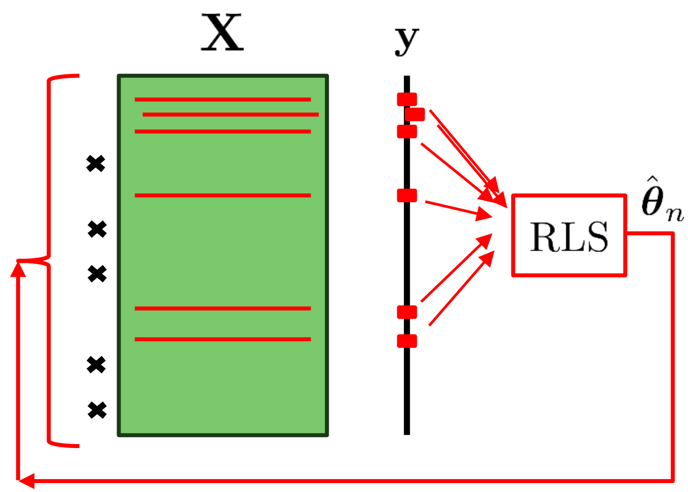
Censoring vis-a-vis random projections

- RPs for linear regressions [Mahoney'11], [Woodruff'14]
  - Data-agnostic reduction; preconditioning costs $\mathcal{O}(pD\log D)$



$$\hat{\boldsymbol{\theta}}_d = \arg\min_{\boldsymbol{\theta}} \|\mathbf{S}_d\mathbf{HB}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\|_2^2$$

- AC for linear regressions
  - Data-driven measurement selection
  - Suitable also for streaming data
  - Minimal memory requirements

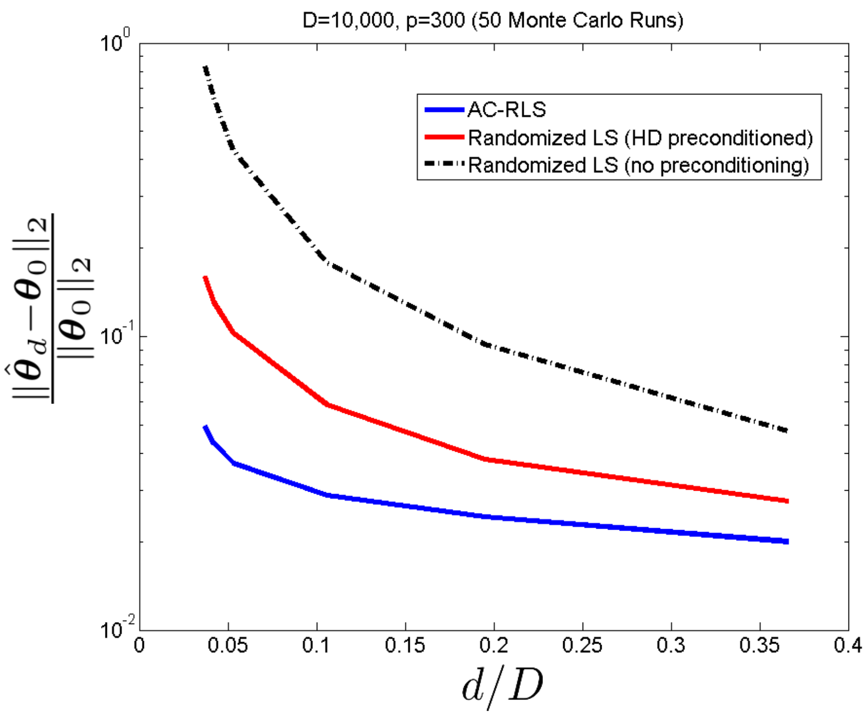


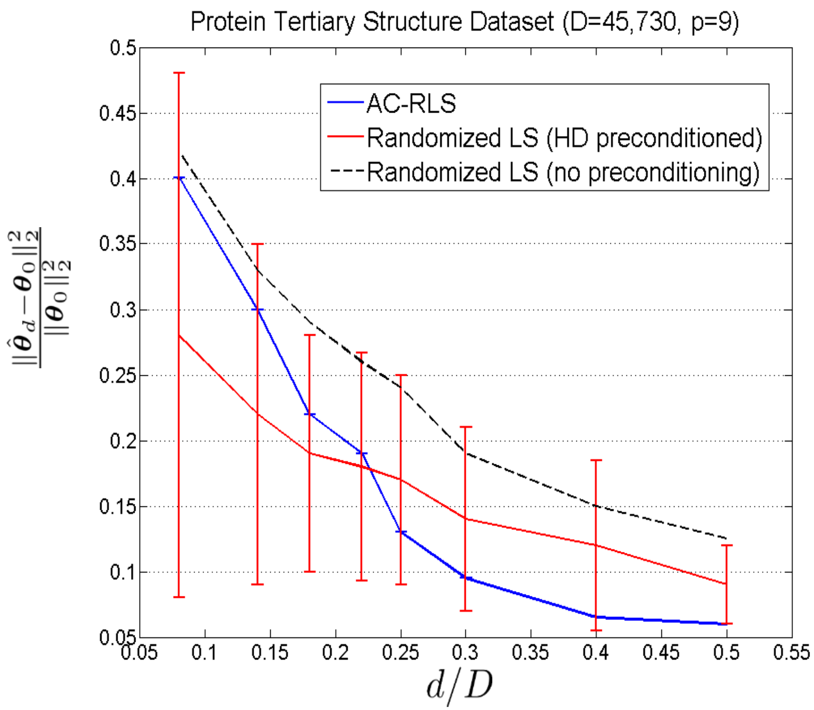
- AC interpretations
  - Reveals `causal' support vectors
  - Censors data with low LLRs: $\log[p(y_n;\boldsymbol{\theta}_o)\ /\ p(y_n;\boldsymbol{\theta}_{n-1})] < \tau$

Performance comparison

- **Synthetic**: *D=10,000, p=300* (50 MC runs); **Real data**: $\boldsymbol{\theta}_0, \sigma$ estimated from full set

Highly non-uniform data





- AC-RLS outperforms alternatives at comparable complexity
- Robust to uniform (all ``important") rows of ***X*** ; **Q:** Time-varying parameters**?**

D. K. Berberidis and G. B. Giannakis, "Data Sketching for Large-Scale Kalman Filtering," *IEEE Trans. on Signal Processing*, vol. 65, pp. 3688 - 3701, Aug. 2017.

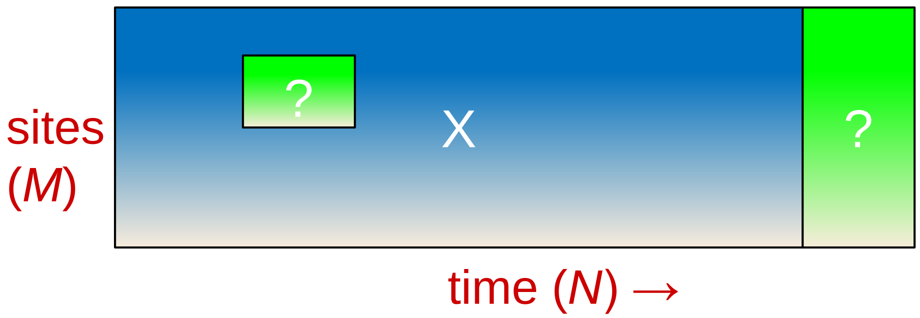
Spatio-temporal load forecasting

- Essential for economic operation of power systems
  - Economic dispatch, OPF, unit commitment (~hour)
  - Reliability assurance and hydrothermal coordination (~week)
  - Strategic generation and transmission planning (~year)

- **Prior art:** time-series models (ARMA/ARIMA/ARIMAX) [Shahidehpour et al'02]

- **Challenges:** account for spatiotemporal patterns; load volatility due to EVs

- **Problem:** given load measurements at M sites and N time slots
  - Predict load at sites/times that data are unavailable; impute past; forecast future demand



Low-rank plus sparse non-negative factors

- Load matrix obeys low-rank plus sparse non-negative bi-factor model

$$\boxed{\mathbf{X} \sim \mathbf{L} + \mathbf{A}\mathbf{B}^T}$$

  - Low-rank ***L*** due to periodicities (daily, weekly, monthly), and latent factors (user preference, temperature)
  - Non-negative matrix factorization $\boldsymbol{AB}^{\mathrm{T}}$ captures load clusters

- Identifiability issues

  - **X** ~ **L** + **S** (**L**: low-rank; **S**: sparse) [Candes et al'11], [Wright'13]
  - **X** ~ **L** + **CS** (**L**: low-rank; **S**: sparse; **C**: given) [Mardani et al'13]
  - Identifiability of our model is plausible but yet to be established

$$\min_{\mathbf{L},\mathbf{A}\geq 0,\mathbf{B}\geq 0} \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X}-\mathbf{L}-\mathbf{A}\mathbf{B}^T)\|_F^2 + \lambda\|\mathbf{L}\|_* + \mu_1(\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1)$$

Load inference algorithm

- Equivalent formulation $\mathbf{P} \in \mathbb{R}^{M\times r}, \mathbf{Q} \in \mathbb{R}^{N\times r}, \mathrm{rank}(L) \leq r$

$$\min_{\mathbf{P},\mathbf{Q},\mathbf{A}\geq 0,\mathbf{B}\geq 0} \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X}-\mathbf{P}\mathbf{Q}^T-\mathbf{A}\mathbf{B}^T)\|_F^2 + \frac{\lambda}{2}(\|\mathbf{P}\|_F^2+\|\mathbf{Q}\|_F^2) + \mu_1(\|\mathbf{A}\|_1+\|\mathbf{B}\|_1),$$

- Solved via block coordinate descent; closed-form per iteration
- Kernelized formulation allows extrapolation [Bazerque-GG'13]

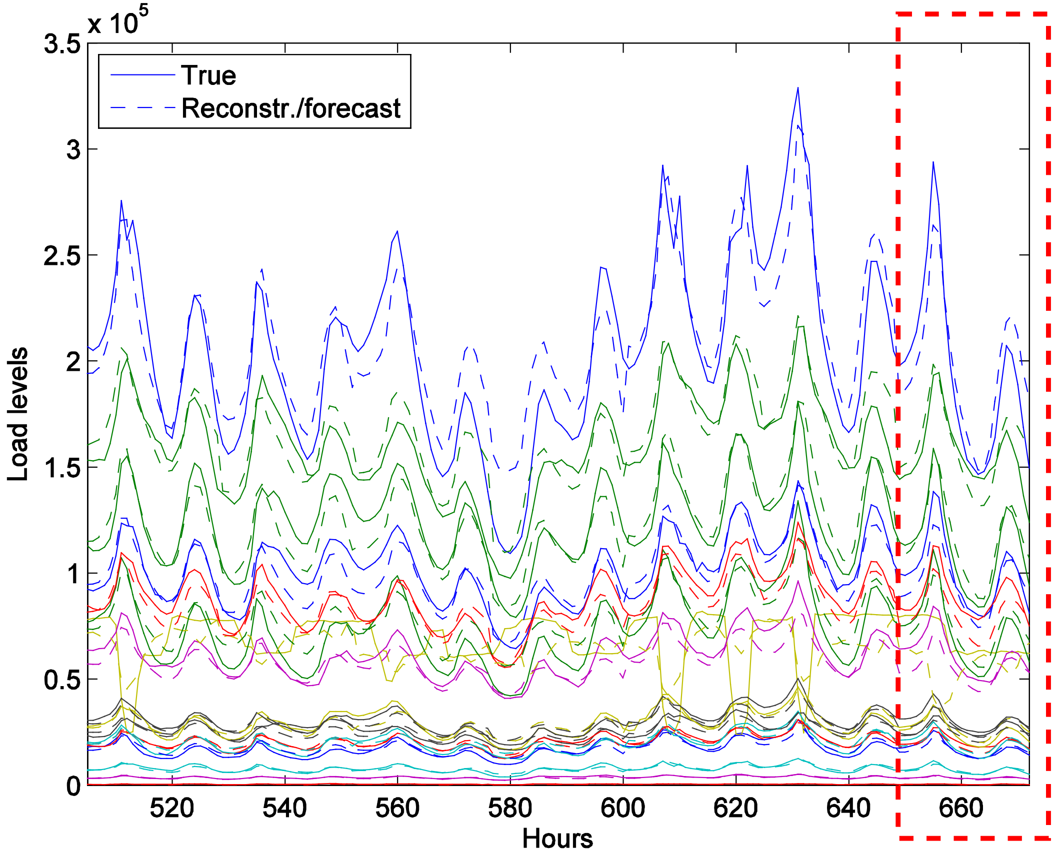
$$\begin{aligned}\min_{\mathbf{P},\mathbf{Q},\mathbf{A}\geq 0,\mathbf{B}\geq 0}\ & \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X}-\mathbf{P}\mathbf{Q}^T-\mathbf{A}\mathbf{B}^T)\|_F^2 \\ & + \frac{\lambda}{2}\left[\mathrm{tr}(\mathbf{P}^T\mathbf{R}_p^{-1}\mathbf{P}) + \mathrm{tr}(\mathbf{Q}^T\mathbf{R}_q^{-1}\mathbf{Q})\right] \\ & + \mu_1(\|\mathbf{A}\|_1+\|\mathbf{B}\|_1) + \frac{\mu_2}{2}\left[\mathrm{tr}(\mathbf{A}^T\mathbf{R}_a^{-1}\mathbf{A}) + \mathrm{tr}(\mathbf{B}^T\mathbf{R}_b^{-1}\mathbf{B})\right]\end{aligned}$$

- $\boldsymbol{R}_p$, $\boldsymbol{R}_q$, $\boldsymbol{R}_a$, $\boldsymbol{R}_b$: positive-definite sample covariances (kernels)

J. A. Bazerque and G. B. Giannakis, "Nonparametric Basis Pursuit via Sparse Kernel-based Learning," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 112-125, July 2013.

Test with real data

- M = 17 sites
- N = 4 * 7 * 24 (4 weeks)
- r = ρ = 5
- Forecast the last 24 hou
  (RMSE = 0.1)



solid: true
dashed: forecast

S.-J. Kim and G. B. Giannakis, ``Forecasting Loads and Renewables via Low Rank and Sparse Matrix Factorization," *Proc. of Asilomar Conf. on Signals, Systems, and Computers*, Nov. 3-6, 2013.

Roadmap

- Context and motivation
- Estimation with big data
- Large-scale data and graph clustering
  - Sketching and validation
  - Spectral clustering
  - Sketched subspace clustering
- Closing comments

Big data clustering

- **Clustering**: Given $\{\mathbf{x}_n\}_{n=1}^{N}$, or their distances, assign them to *K* clusters

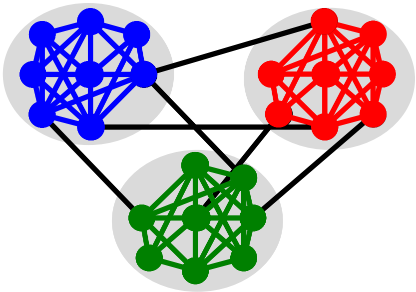
$$\boxed{\begin{aligned}&\min_{\mathbf{C},\boldsymbol{\Pi}} \sum_n \|\boldsymbol{x}_n - \mathbf{C}\boldsymbol{\pi}_n\|_2^2 \\ &\text{s.to } \mathbf{1}^\top \boldsymbol{\pi}_n = 1,\ \boldsymbol{\pi}_n \succeq \mathbf{0},\ n = 1, \ldots N\end{aligned}}$$

$\mathbf{C} := [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K]$

Centroids

$\boldsymbol{\Pi} := [\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n]$

Assignments



- **Hard clustering:** $\boldsymbol{\pi}_n \in \{0,1\}^K$ NP-hard!
- **Soft clustering:** $\boldsymbol{\pi}_n \in [0,1]^K$

- **AEG context:** consumer profiling, controlled islanding

- **K-means:** locally optimal, but simple; complexity O(*NDKI*)

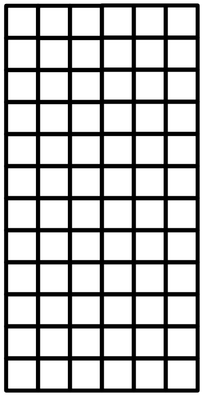
Q. What if $\boxed{N \gg}$ and/or $\boxed{D \gg}$ **?**

A1. Random Projections: Use *dxD* matrix ***R*** to form ***RX***; apply *K*-means in *d*-space

C. Boutsidis, A. Zousias, P. Drineas, and M. W. Mahoney, “Randomized dimensionality reduction for K-means clustering,” *IEEE Trans. on Information Theory*, vol. 61, pp. 1045-1062, Feb. 2015.

Random sketching and validation (SkeVa)

- Randomly select $d \ll D$ "informative" dimensions

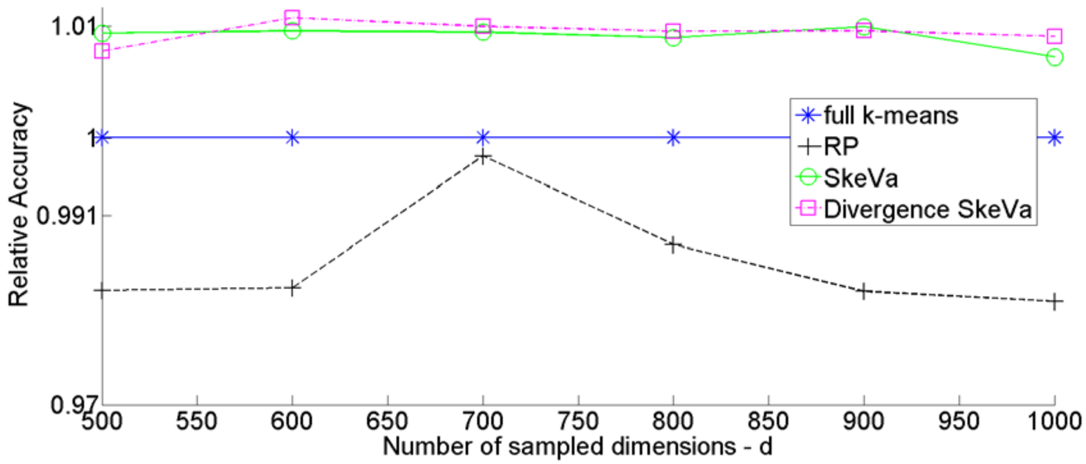


- Algorithm For $r = 1, \ldots, R_{\max}$
  - Sketch $d \ll D$ dimensions: $\mathbf{X} \rightarrow \check{\mathbf{X}}^{(r)} \in \mathbb{R}^{d \times N}$
  - Run k-means on $\check{\mathbf{X}}^{(r)} \rightarrow \{\check{\mathcal{C}}_k^{(r)}\}_{k=1}^K, \{\check{\boldsymbol{c}}_k^{(r)}\}_{k=1}^K$
  - Re-sketch $d' \leq D - d$ dimensions $\rightarrow \check{\mathbf{X}}^{(r')} \in \mathbb{R}^{d' \times N}$
  - Augment centroids $\bar{\boldsymbol{c}}_k^{(r)} := [\check{\boldsymbol{c}}_k^{(r)\top}, \check{\boldsymbol{c}}_k^{(r')\top}]^\top \quad \forall k,\ \check{\boldsymbol{c}}_k^{(r')} = \frac{1}{|\check{\mathcal{C}}_k^{(r)}|} \sum_{\check{\boldsymbol{x}}_n^{(r)} \in \check{\mathcal{C}}_k^{(r)}} \check{\boldsymbol{x}}_n^{(r')}$
  - Validate using consensus set $\mathcal{S}^{(r)} = \{\boldsymbol{x}_n | \check{\boldsymbol{x}}_n^r \in \check{\mathcal{C}}_{k_1}^{(r)}, \bar{\boldsymbol{x}}_n^r \in \bar{\mathcal{C}}_{k_2}^{(r)}, \quad \text{and} \quad k_1 = k_2\}$
    - $r^* = \arg\max_r f(\mathcal{S}^{(r)})$

- Similar approaches possible for $N \gg$
- Sequential and kernel variants available

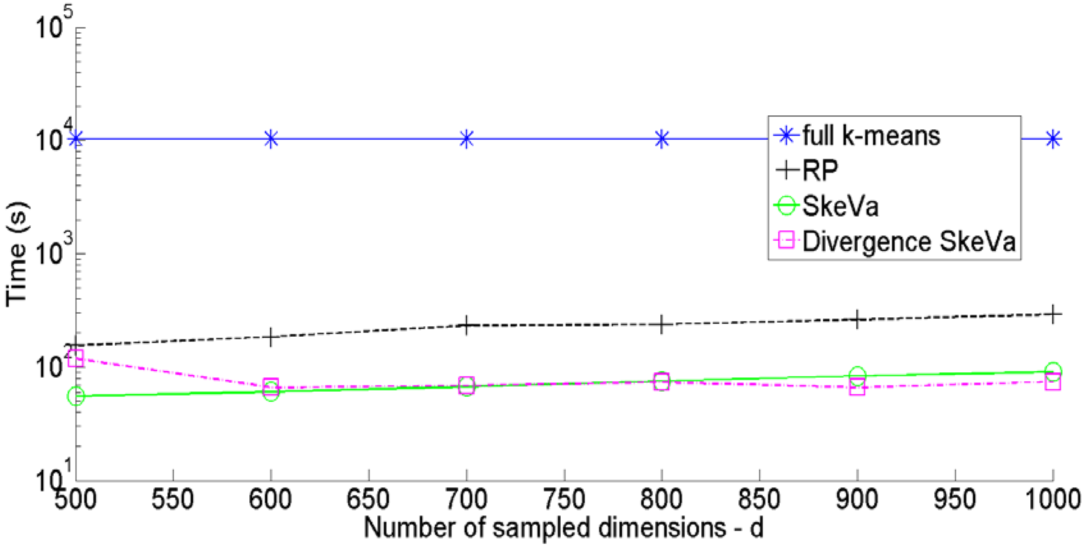
P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Sketch and Validate for Big Data Clustering," *IEEE Journal on Special Topics in Signal Processing,* vol. 9, pp. 678-690, June 2015.

RP versus SkeVa comparisons



KDDb dataset (subset)

D **= 2,990,384,** ***N*** **= 10,000,** *K* = 2



RP: [Boutsidis etal '15]

versus SkeVa

Performance and SkeVa generalizations

- Di-SkeVa fully parallelizable

Q. How many samples/draws SkeVa needs**?**

A. For independent draws, $R_{\max}$ can be lower bounded

Proposition 2. For a given probability π_s of a successful Di-SkeVa draw *r* quantified by pdf dist. *Δ,* the number of draws is lower bounded w.h.p. *q* by

$$R_{\max} \geq \frac{\log(1-\pi_s)}{\log\left(1-\Delta_0^{-1} E[\Delta(p_0,\hat{p})]\right)}$$

- Bound can be estimated online

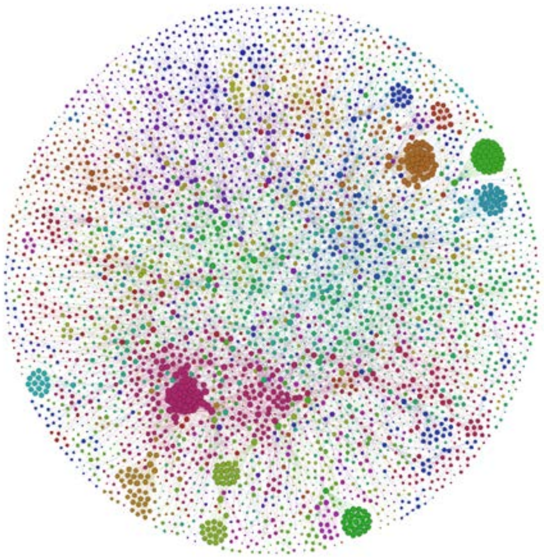
$$\bar{\Delta}^{(r)}(p_0,\hat{p}) = \frac{1}{r}\sum_{i=1}^{r}\Delta(p_0^{(i)},\hat{p}^{(i)}) \qquad \hat{\Delta}_0^{(r)} = \left(\sqrt{-\frac{2\log(q/2)}{n\sigma_\kappa(4\pi)^{D/2}}} + \bar{\Delta}^{(r)}(\tilde{p},\hat{p}) + \bar{\Delta}^{(r)}(\tilde{p},p_0)\right)^2$$

- SkeVa module can be used for **spectral clustering** and **subspace clustering**

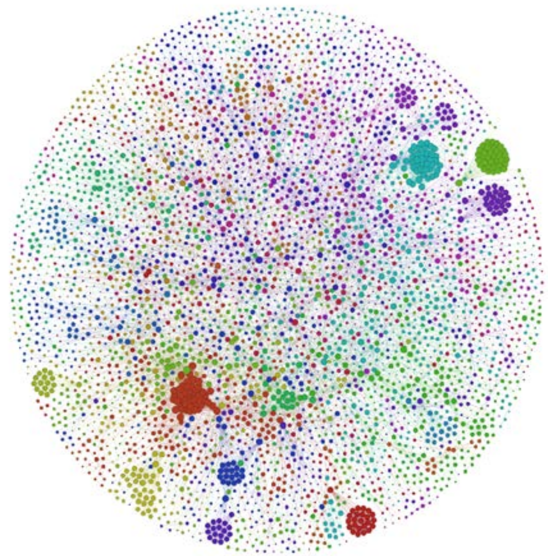
Scalable clustering of cellular blocks

- Kernel K-means instrumental for partitioning of large graphs (spectral clustering)
  - Relies on graph Laplacian to capture nodal correlations

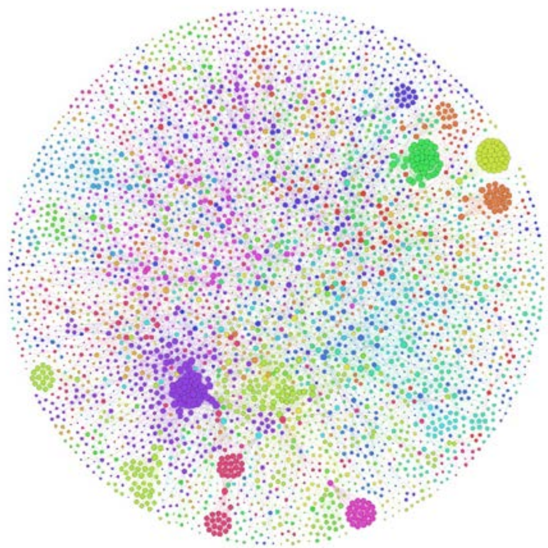
arXiv collaboration network (General Relativity): *N=4,158* nodes, 13,422 edges, *K = 36* [Leskovec'11]



Spectral clustering
3.1 sec



SkeVa (*n = 500)*
0.5 sec



SkeVa (*n=1,000*)
0.85 sec

- For $D \gg$, kernel-based SkeVa reduces complexity to $\mathcal{O}(d)$

P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Spectral clustering of large-scale communities via random sketching and validation," *Proc. Conf. on Info. Science and Systems,* Baltimore, Maryland, March 18-20, 2015.

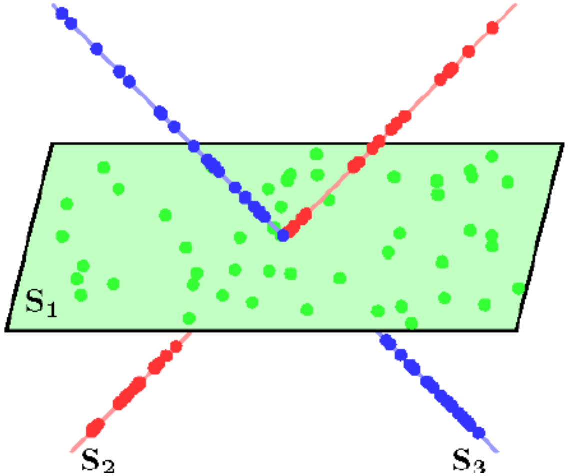
Subspace clustering

- Given high-dimensional data $\mathbf{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] : D\text{-by-}N$
  - Find *K* subspaces (clusters) $\{\mathcal{S}_i\}_{i=1}^K$,
  - their dimensions $\{d_i = \dim(\mathcal{S}_i)\}_{i=1}^K$
  - their centroids $\{\boldsymbol{\mu}_i\}_{i=1}^K$
  - their bases $\{\mathbf{U}_i\}_{i=1}^K$
  - their low-dimensional representations $\{\boldsymbol{y}_j \in \mathbb{R}^{\{d_i\}}\}_{j=1}^N$

$$\mathcal{S}_i = \{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} = \boldsymbol{\mu}_i + \mathbf{U}_i \boldsymbol{y}\}$$

$$\min_{\{\boldsymbol{\mu}_i\}\{\mathbf{U}_i\}\{\boldsymbol{y}_i\}\{\pi_{ij}\}} \quad \sum_{i=1}^K \sum_{j=1}^N \pi_{ij} \|\boldsymbol{x}_j - \boldsymbol{\mu}_i - \mathbf{U}_i \boldsymbol{y}_j\|^2$$

$$\text{subject to} \quad \pi_{ij} \in \{0,1\} \text{ and } \sum_{i=1}^K \pi_{ij} = 1$$



- Encapsulates *K*-means and PCA

R. Vidal, "A tutorial on subspace clustering," *IEEE Signal Processing Magazine,* pp. 52-68, 2010.

State-of-the-art batch approaches

- A point in *d*-dim subspace as a lin. comb. of *d*(*d+1*) points in the same space

$$\boldsymbol{x}_i \in S_k \rightarrow \boldsymbol{x}_i = \sum_{j:\boldsymbol{x}_j \in S_k} z_{ij}\boldsymbol{x}_j$$

- Sparse subspace clustering (SSC): Relies on sparsity to choose nearest neighbors

$$\boxed{\begin{aligned} &\min_{\mathbf{Z}} \quad \|\mathbf{Z}\|_1 + \frac{\lambda}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 \\ &\text{s.to} \quad \mathbf{Z}\mathbf{1} = \mathbf{1}; \quad \text{diag}(\mathbf{Z}) = \mathbf{0} \end{aligned}}$$

- Low-rank representation (LRR): Low-rank instead

$$\boxed{\min_{\mathbf{Z}} \quad \|\mathbf{Z}\|_* + \frac{\lambda}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2}$$

- Least-squares regression (LSR): Frobenius norm instead

$$\boxed{\min_{\mathbf{Z}} \quad \|\mathbf{Z}\|_F^2 + \frac{\lambda}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2}$$

Computationally heavy for large *N*

  - Use spectral clustering with affinity matrix: $\mathbf{W} = |\mathbf{Z}| + |\mathbf{Z}^T|$

E. Elhamifar, and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications,“ *IEEE Trans. on PAMI*, pp. 2765 -2781, 2013.

Sketched subspace clustering

- Use a smaller *D-by-n* "basis": $\mathbf{B}$

Solve: $\min_{\mathbf{Z}} \quad \|\mathbf{Z}\| + \frac{\lambda}{2}\|\mathbf{X} - \mathbf{BZ}\|_F^2$ $\quad \mathcal{O}(nN)$ variables to optimize

- Use spectral clustering on $\mathbf{Z}$

Q. How to select **B?** **A.** $\mathbf{B} = \mathbf{XR} \quad \mathbf{R} : N \times n$

$$[\mathbf{R}]_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p); \mathcal{N}(0,1); \text{Rademacher}(1/2)$$

Prop. 3a If ***R*** is *Nxn* JLT, $n > \text{rank}(\mathbf{X}) = \rho$, then $\text{range}(\mathbf{X}) = \text{range}(\mathbf{B})$ whp

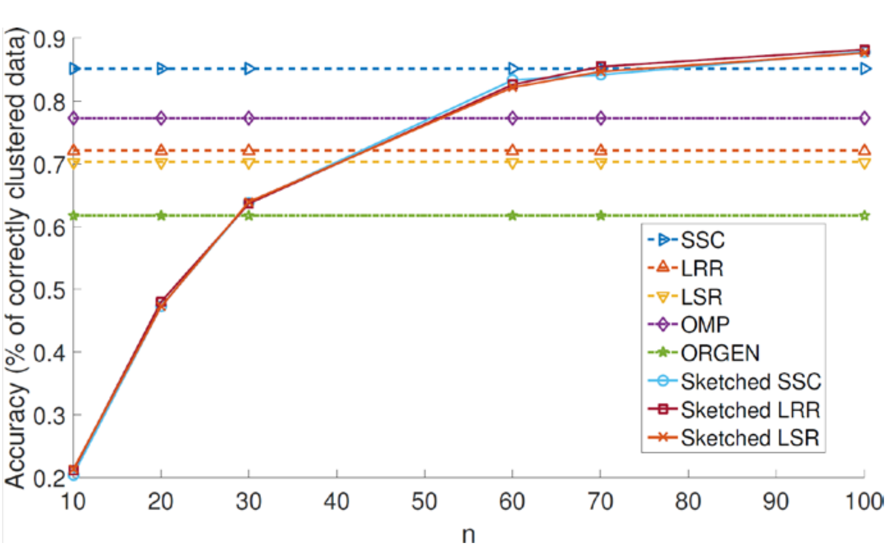
Prop. 3b If $f^*(\boldsymbol{x}_i) = \mathbf{X}\boldsymbol{z}_i, \hat{f}(\boldsymbol{x}_i) = \mathbf{XR}\boldsymbol{z}_i$ with ***R*** *Nxn* JLT, then

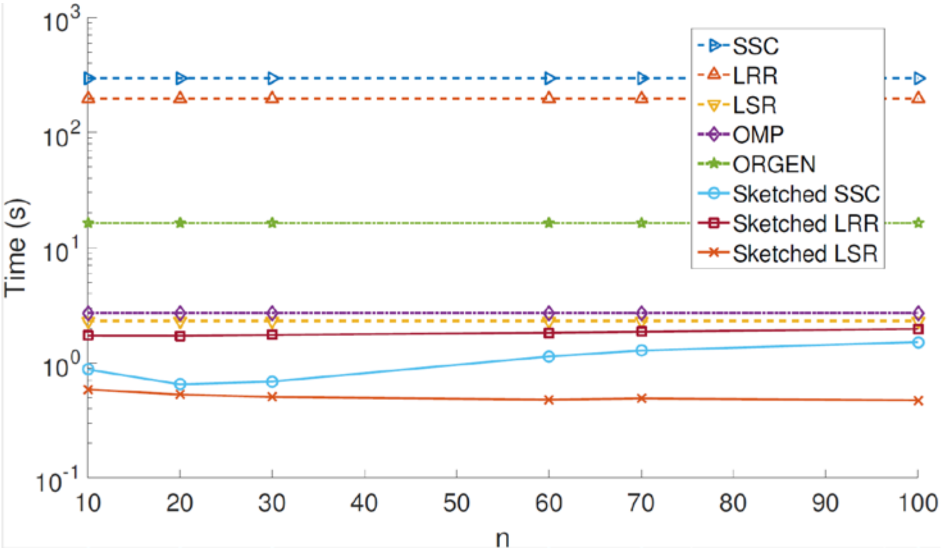
$$\|f^*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\|_2 \leq c_1\lambda\sigma_\ell^2 + c_2$$

P. A. Traganitis and G. B. Giannakis, "Sketched subspace clustering," arXiv:1707.07196, 2017.

Extended Yale Face database B

- ❑ ***N = 2,048, D = 2,016, K = 10***

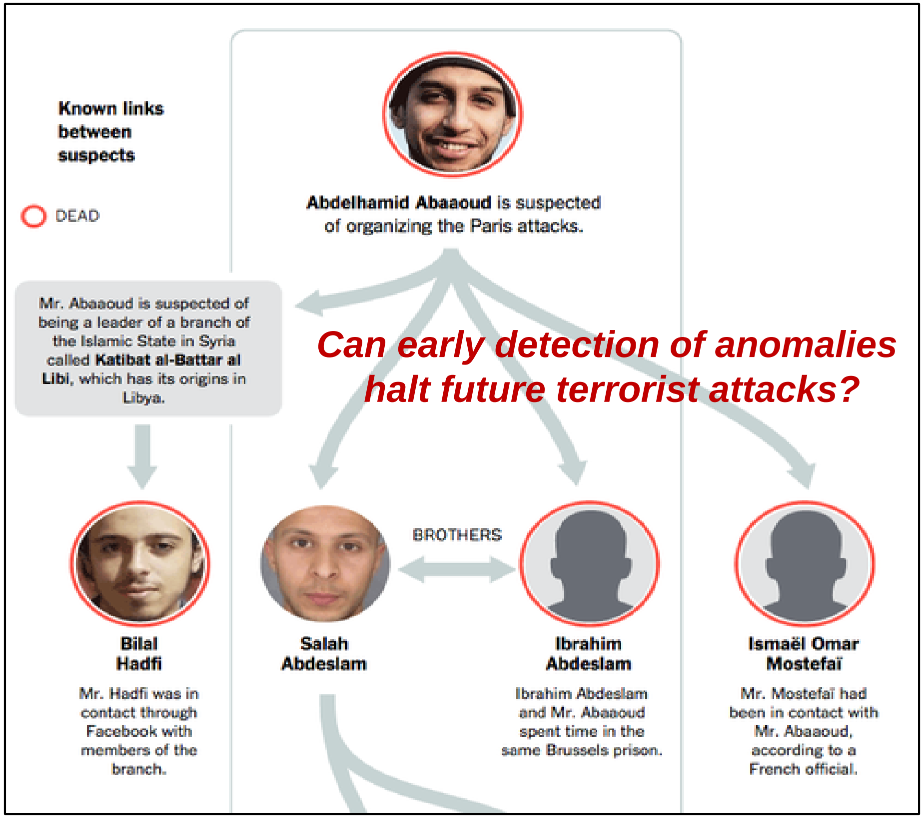




A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *IEEE TPAMI*, vol. 23, no. 6, pp. 643–660, 2001.

Anomalies in social (or AEG) graphs

- To identify e.g., "strange" users and "atypical" behavior



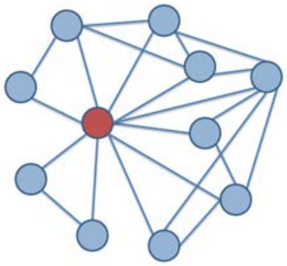
- **Examples**
  - E-mail spammers
  - Cybercriminals
  - Terrorist cells

- **Egonet features**
  - Degree, number of edges, centrality, betweeness, …

- **Challenge:** Too many users, BUT few features per user
- **Approach:** Adopt **"egonet"** features, and leverage structure; e.g., sparsity and low rank

B. Baingana, P. Traganitis, G. Mateos, and G. B. Giannakis, "Big data analytics for social networks," *Graph Analysis for Social Media*, I. Pitas, Editor, CRC Press, 2015.

Low-rank plus sparse model



- Egonets can unveil anomalous behavior [Akoglu et al'10]
- *N*-node graph with egonet features $\mathbf{Y} := [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{D \times N}$
  - $\mathbf{y}_n := [y_{n,1}, \ldots, y_{n,D}]^\top$ collects *D* features for egonet *n*
  - Nominal features related via **"power law"** while anomalies are **sparse**

$$\underbrace{\mathbf{Y} = \mathbf{X}}_{\text{Low-rank nominal features}} + \underbrace{\mathbf{O}}_{\text{Sparse outlier matrix}} + \mathbf{E}$$

- Account for **"misses"** via sampling operator $\mathcal{P}_\Omega$

$$\boxed{\mathcal{P}_\Omega(\mathbf{Y}) = \mathcal{P}_\Omega(\mathbf{X} + \mathbf{O} + \mathbf{E})}$$

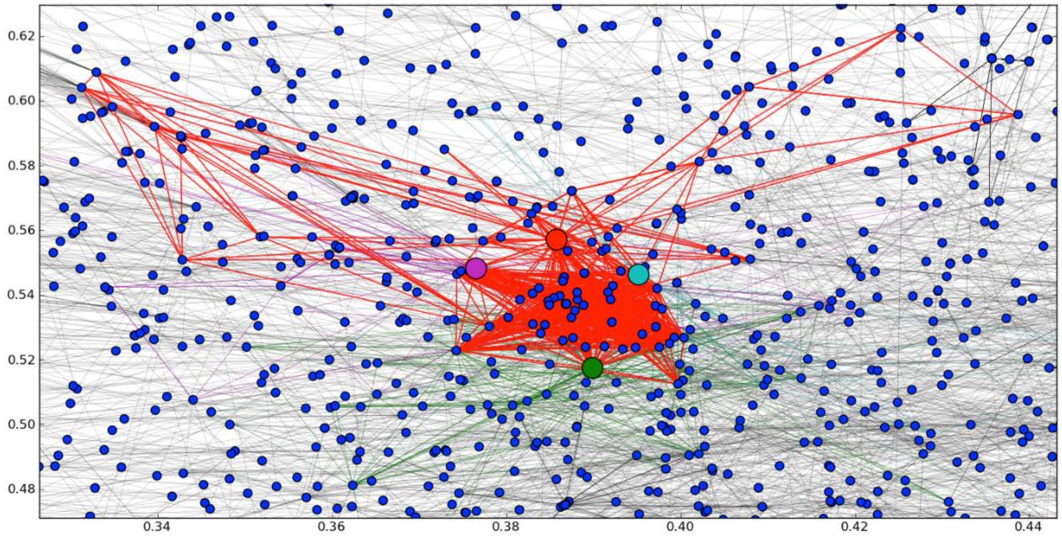
Robust low-rank component pursuit

- Low-rank- plus sparsity-promoting estimator

$$\min_{\{\mathbf{X},\mathbf{O}\}} \|\mathcal{P}_\Omega(\mathbf{Y}-\mathbf{X}-\mathbf{O})\|_F^2 + \lambda_*\|\mathbf{X}\|_* + \lambda_1\|\mathbf{O}\|_1$$

  - $\|\mathbf{O}\|_1 := \sum_{d,n} |o_{d,n}|$ and $\|\mathbf{X}\|_* := \sum_i \sigma_i(\mathbf{X})$

- **Numerical test:** Anomalies in ***ArXiv*** collaboration network (General Relativity co-authors)



  - D = 9, N = 5,242 nodes
  - Observed Jan. '93 – Apr.'03

M. Mardani, G. Mateos, and G. B. Giannakis, ``Recovery of low rank plus compressed sparse matrices with application to unveiling traffic anomalies,'' *IEEE Trans. Info. Theory*, vol. 59, no. 8, pp. 5186-5205, Aug. 2013.

Closing comments

- **Large-scale estimation and scalable clustering**
  - Regression and tracking dynamic data
  - Nonlinear non-parametric function approximation
  - Clustering massive, high-dimensional data and graphs
- **Other key Big Energy Data tasks**
  - Visualization, mining, dynamics, privacy, and security
- **Enabling tools for Big Data**
  - Acquisition, communication, processing, and storage
  - Fundamental theory, performance analysis
    decentralized, robust, large-scale, and parallel optimization
  - Scalable computing platforms

Thank You!

K. Slavakis, G. B. Giannakis, and G. Mateos, "Modeling and optimization for Big Data analytics," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 18-31, Sep. 2014.