ACCESS Data Analytics Workshop
Stockholm, Sweden, May 15, 2017

KTH VETENSKAP OCH KONST

Adaptive Sketching and Validation for Learning from Large-Scale Data

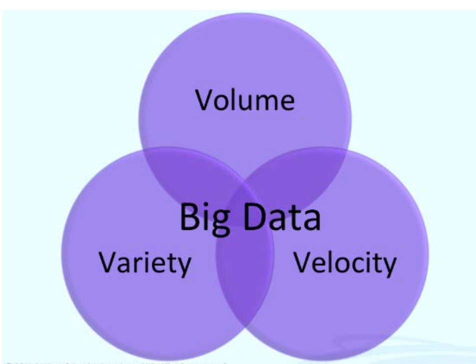
Georgios B. Giannakis

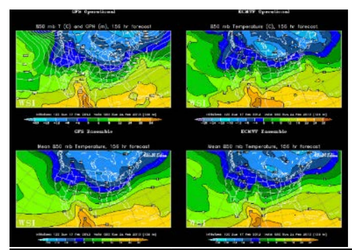
Acknowledgments: D. Berberidis, P. Traganitis; Dr. M. Mardani, and Prof. G. Mateos
NSF 1343860, 1442686, 1514056, NSF-AFOSR 1500713,
MURI-FA9550-10-1-0567, and NIH 1R01GM104975-01

UNIVERSITY OF MINNESOTA
Driven to Discover℠

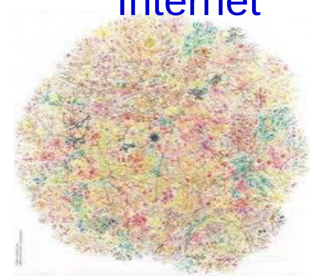
Learning from “Big Data”

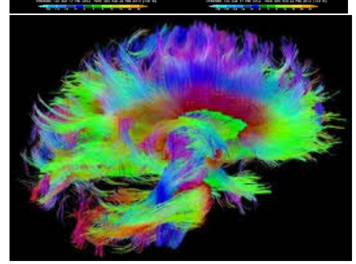
- Challenges
  - Big size ($D \gg$ and/or $N \gg$)
  - Fast streaming
  - Incomplete
  - Noise and outliers
- Opportunities in key tasks
  - Dimensionality reduction
  - Online and robust regression, classification and clustering
  - Denoising and imputation





Internet





Roadmap

- Context and motivation
- Large-scale linear regressions
  - Random projections for data sketching
  - Adaptive censoring of uninformative data
- Large-scale data and graph clustering
- Leveraging sparsity and low rank for anomalies and tensors
- Closing comments

Random projections for data sketching

Ordinary least-squares (LS) Given $\mathbf{y} \in \mathbb{R}^D, \ \mathbf{X} \in \mathbb{R}^{D \times p}$

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

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

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

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

- 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]

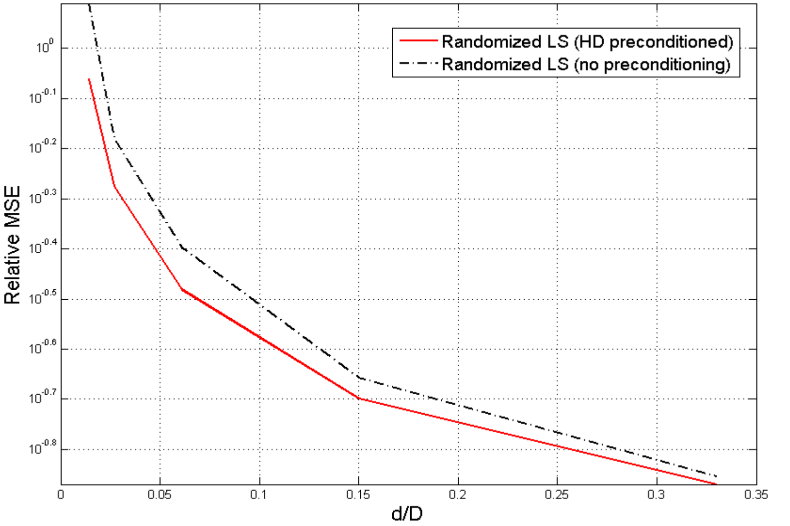
Theorem. For any $\epsilon > 0$, if $d = \mathcal{O}(p \log p/\epsilon^2)$ then w.h.p.

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

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

$\kappa(\mathbf{X})$ 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.

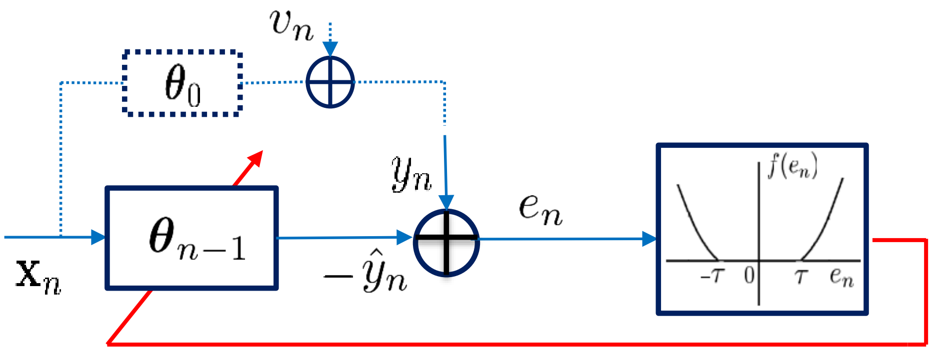
Online censoring for large-scale regressions

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

- Adaptive censoring (AC) rule:

  Censor if

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



- Criterion

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

- 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)

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

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

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

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

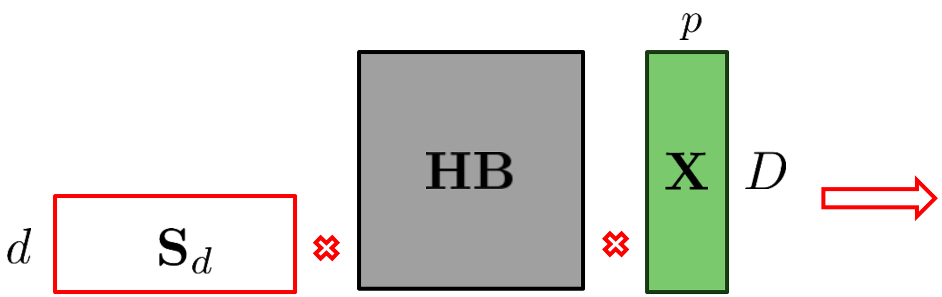
Proposition 1 **AC-RLS** $\frac{1}{n}\mathrm{tr}\left(\mathbf{R}_x^{-1}\right)\sigma^2 \leq \mathbf{E}\left[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{1}{n}\frac{\mathrm{tr}\left(\mathbf{R}_x^{-1}\right)\sigma^2}{2Q(\tau)} \quad \forall n \geq k$

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

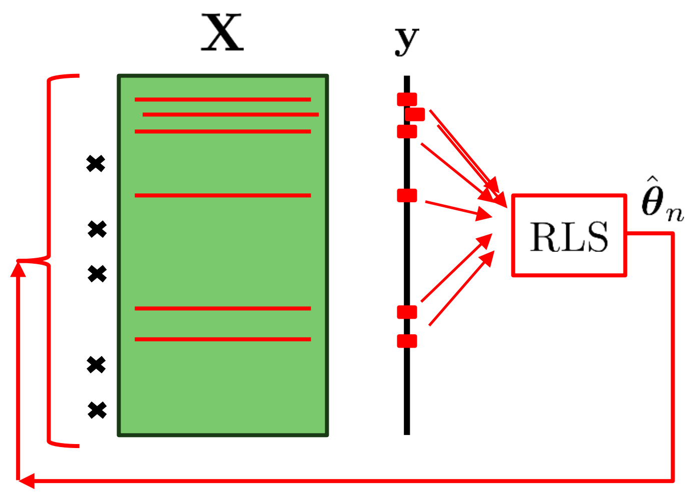
D. K. Berberidis, V. Kekatos, and G. B. Giannakis, "Online Censoring for Large-Scale Regressions with Application to Streaming Big Data," *IEEE Trans. on Signal Processing*, vol. 64, pp. 3854-3867, Aug. 2016.

Censoring vis-a-vis random projections

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



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

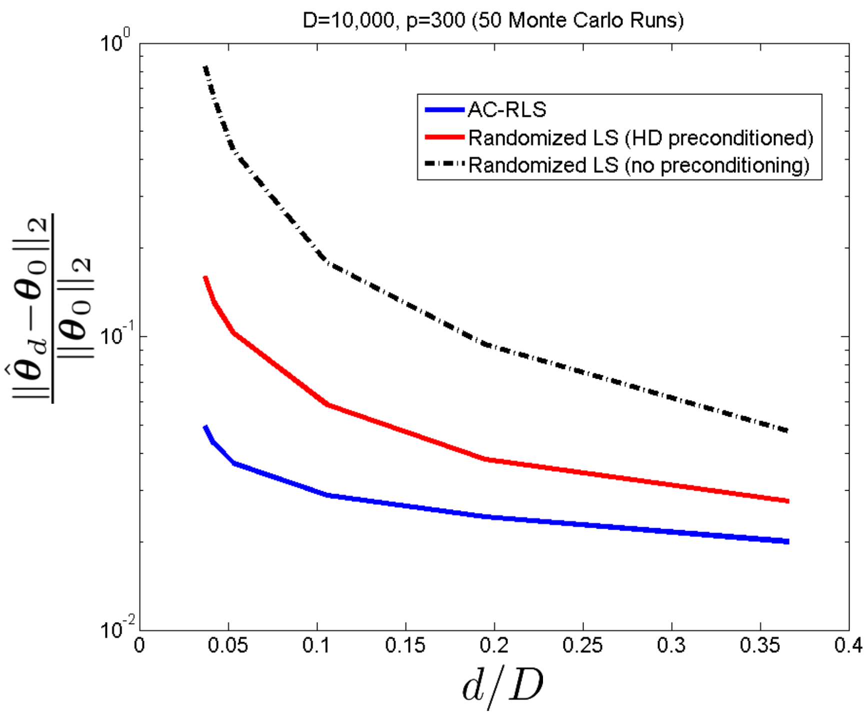


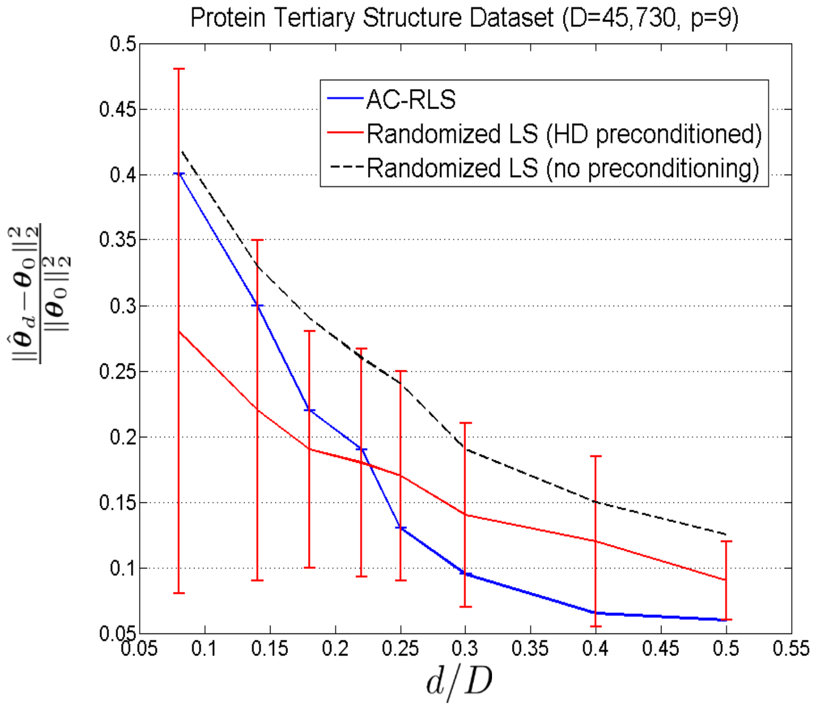
- AC for linear regressions
  - Data-driven measurement selection
  - Suitable also for streaming data
  - Minimal memory requirements
- AC interpretations
  - Reveals 'causal' support vectors
  - Censors data with low LLRs: $\log[p(y_n; \boldsymbol{\theta}_o) \;/\; p(y_n; \boldsymbol{\theta}_{n-1})] < \tau$

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 $\boldsymbol{X}$; **Q:** Time-varying parameters**?**

Roadmap

- Context and motivation
- Large-scale linear regressions
- Large-scale data and graph clustering
  - Random sketching and validation (SkeVa)
  - SkeVa-based spectral and subspace clustering
- Leveraging sparsity and low rank for anomalies and tensors
- Closing comments

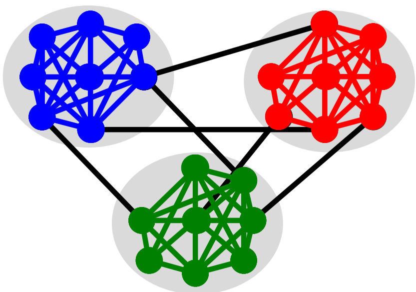
Big data clustering

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

$$\min_{\mathbf{C},\mathbf{\Pi}} \sum_n \|\boldsymbol{x}_n - \mathbf{C}\boldsymbol{\pi}_n\|_2^2 + \lambda \|\boldsymbol{\pi}_n\|_1$$
$$\text{s.to } \mathbf{1}^\top \boldsymbol{\pi}_n = 1,\ \boldsymbol{\pi}_n \succeq \mathbf{0},\ n = 1, \ldots N$$

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

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



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

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

- Probabilistic clustering amounts to pdf estimation
  - Gaussian mixtures (EM-based estimation)
  - Regularizer can account for unknown *K*

$$p(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \underbrace{p(\boldsymbol{x}; \boldsymbol{\theta}_k)}_{p(\boldsymbol{x}|\mathcal{C}_k)}$$

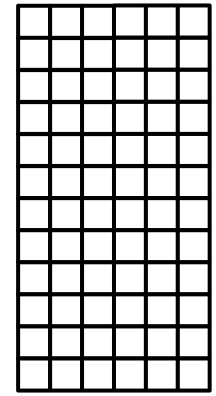
Q. What if $N \gg$ and/or $D \gg$?

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

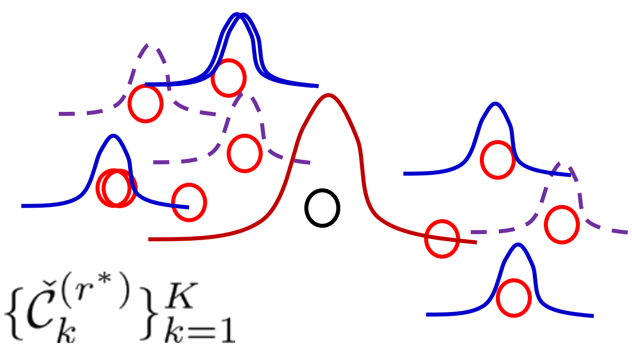
P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Sketch and Validate for Big Data Clustering," *IEEE Journal on Special Topics in Signal Processing,* vol. 9, pp. 678-690, June 2015.

Divergence-based SkeVa

- **Idea:** "Informative" draws yield reliable estimates of multimodal data pdf!
  - Compare pdf estimates $\hat{p}(\mathbf{x}) := \frac{1}{\nu}\sum_{n=1}^{\nu}\kappa(\mathbf{x}_n, \mathbf{x})$ via "distances"
    - **Integrated square-error** (ISE) $\Delta_{ISE}(p_1||p_2) := \int (p_1(\mathbf{x}) - p_2(\mathbf{x}))^2 d\mathbf{x}$

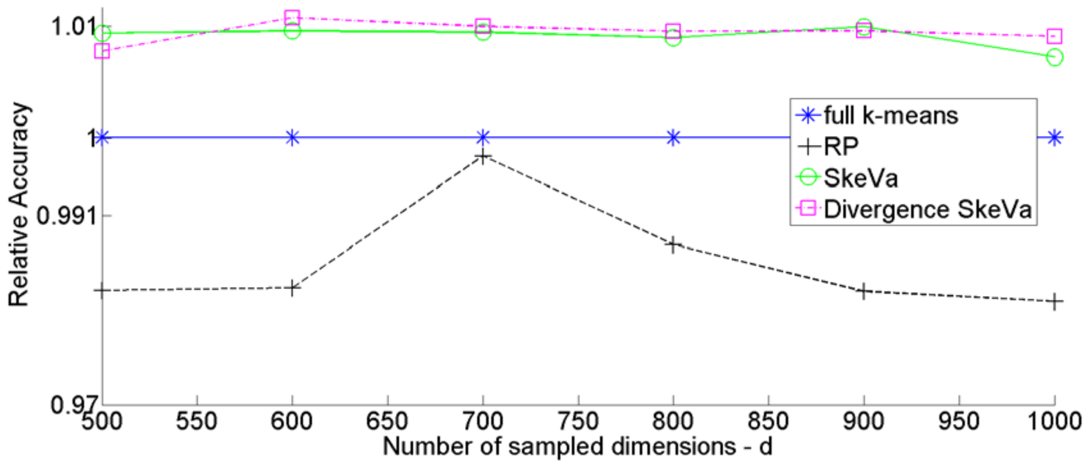
$$\int p_1(\mathbf{x})p_2(\mathbf{x})d\mathbf{x} = \frac{1}{\nu_1\nu_2}\mathbf{1}^\top\mathbf{K}^{(p_1,p_2)}\mathbf{1}$$

- For $r = 1, \ldots, R_{\max}$
  - Sketch ν points $\rightarrow \check{\mathbf{X}}^{(r)} \in \mathbb{R}^{D\times\nu} \rightarrow \check{p}^{(r)}(\mathbf{x}) := \frac{1}{\nu}\sum_n \kappa(\mathbf{x}_n^{(r)}, \mathbf{x})$
  - If $\Delta(\check{p}^{(r)}||\check{p}^0) \geq \Delta_{\max}$, then re-sketch ν' points
  - If $\Delta(\check{p}^{(r)}||\check{p}^{(r')}) \leq \Delta_{\min}$ ✓ $r^* := r$



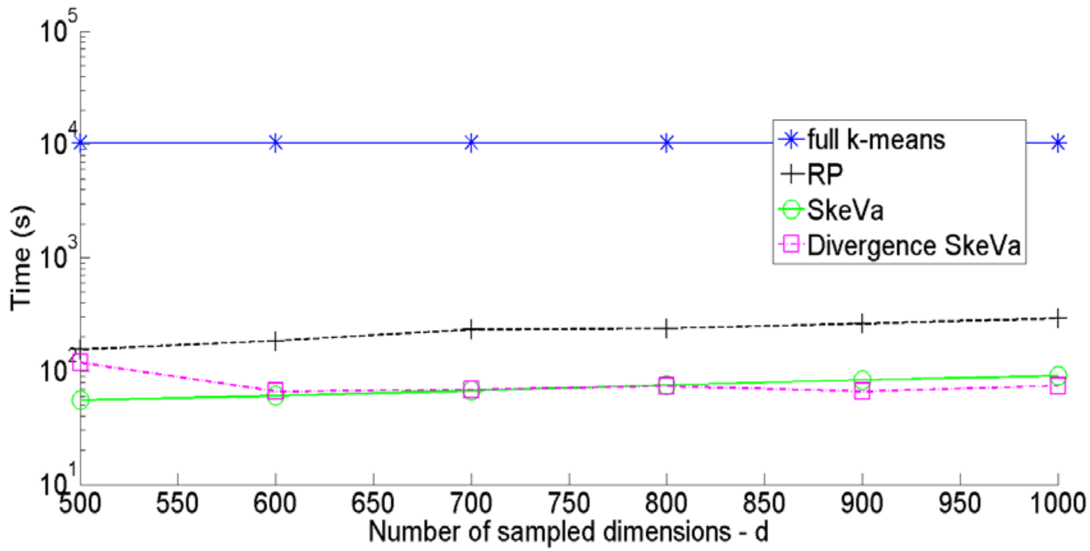
- Cluster $\check{\mathbf{X}}^{(r^*)} \rightarrow \{\check{\mathcal{C}}_k^{(r^*)}\}_{k=1}^K$; associate $\mathbf{X}/\check{\mathbf{X}}^{(r^*)}$ to $\{\check{\mathcal{C}}_k^{(r^*)}\}_{k=1}^K$

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 is 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

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

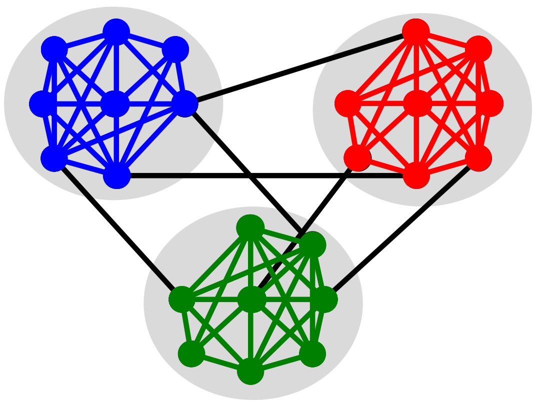
- Bound can be estimated online

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

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

Communities in "big" social nets

- **Community structure** prevalent in "big" networks [Fortunato'10], [Girvan-Newman'02]
  - Strong intra-cluster connections; weak links elsewhere



- Extensively studied problem with many classical tools
  - Graph partitioning [Kernighan et al'70], [Shi et al'00]
  - Modularity maximization [Newman'06]

- "Workhorse" approach: **Spectral Clustering** [Von Luxburg'07]
  - Given weighted adjacency matrix $\mathbf{W}$, want *K* communities

*Compute graph **Laplacian***	→	*Spectral decomposition*	→	***K-means** on rows of K trailing **eigenvectors***
$\mathbf{L} = \mathrm{Diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$		$\mathbf{L} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$		$\mathbf{U}_K := [\mathbf{u}_1, \ldots, \mathbf{u}_K]$

Spectral clustering as kernel K-means

- **Kernel K-means** [Dhillon et al'04]
  - Map data $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$ to higher-dimensional $(\tilde{D} \gg D)$ space $\mathbf{x}_i \rightarrow \boldsymbol{\phi}(\mathbf{x}_i) \in \mathcal{F}$

$$\min_{\{\mathcal{C}_k\}} \sum_{i=1}^{N} \left\| \boldsymbol{\phi}(\mathbf{x}_i) - \frac{1}{|\mathcal{C}_k|} \sum_{j \in \mathcal{C}_k} \boldsymbol{\phi}(\mathbf{x}_j) \right\|^2$$

"kernel trick"

$[\mathbf{K}]_{ij} = \boldsymbol{\phi}^\top(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$

$$\min_{\boldsymbol{\Pi} \in \mathbb{R}^{N \times K}} \operatorname{tr}(\mathbf{K}) - \operatorname{tr}(\boldsymbol{\Pi}^\top \mathbf{K} \boldsymbol{\Pi})$$

  - Assignment matrix: $[\boldsymbol{\Pi}]_{ik} = \begin{cases} \frac{1}{|\mathcal{C}_k|} & \text{if } \mathbf{x}_i \in \mathcal{C}_k \\ 0 & \text{otherwise} \end{cases}$

- Proper kernel choice
  - Kernel K-means $\Longleftrightarrow$ spectral clustering
  - Both rely on similarities $\Longrightarrow$ useful for graph clustering, but do they scale well?

Kernel sketch and validate (K-SkeVa)

- Randomly select $\nu \ll N$ "informative" vertices
- **Algorithm:** For $r = 1, \ldots, R_{\max}$
  - Sketch $\nu \ll N$ vertices: $\mathbf{K} \rightarrow \check{\mathbf{K}}^{(r)} \in \mathbb{R}^{\nu \times \nu}$
  - Run k-means on $\check{\mathbf{K}}^{(r)} \rightarrow \{\check{\mathcal{C}}_k^{(r)}\}_{k=1}^{K}, \{\check{\boldsymbol{\pi}}^{(r)}\}_{k=1}^{K}$
  - Re-sketch $\nu' \leq N - \nu$ vertices $\rightarrow \check{\mathbf{K}}^{(r')} \in \mathbb{R}^{\nu \times (\nu + \nu')}$
  - Re-compute clusters w/ newly sampled ν' vertices $\check{\mathbf{K}}^{(r')} \in \mathbb{R}^{\nu \times (\nu + \nu')} \rightarrow \{\bar{\mathcal{C}}_k^{(r)}\}_{k=1}^{K}$
  - Validate using consensus set $\mathcal{S}^{(r)} = \{\mathbf{x}_n^{(r)} \in \check{\mathbf{X}}^{(r)} \mid \exists k \text{ s.t. } \mathbf{x}_n^{(r)} \in (\check{\mathcal{C}}_k^{(r)} \cap \bar{\mathcal{C}}_k^{(r)})\}$

$$r^* = \arg\min_{r} f(\mathcal{S}^{(r)})$$

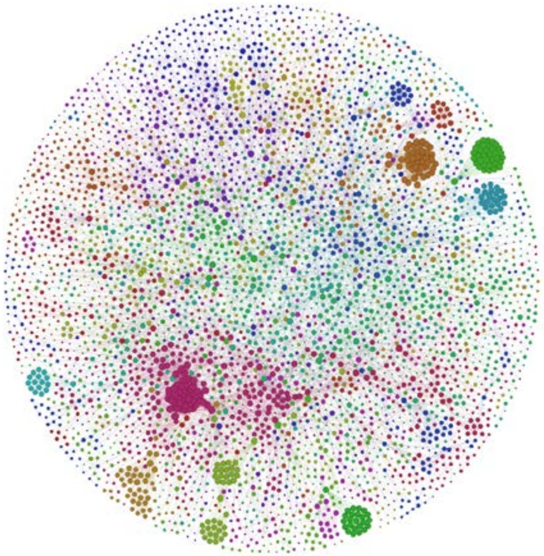
- Fully parallelizable!

P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Spectral clustering of large-scale communities via random sketching and validation," *Proc. of Conf. on Information. Sciences and Systems*, Baltimore, MD, Mar. 2015

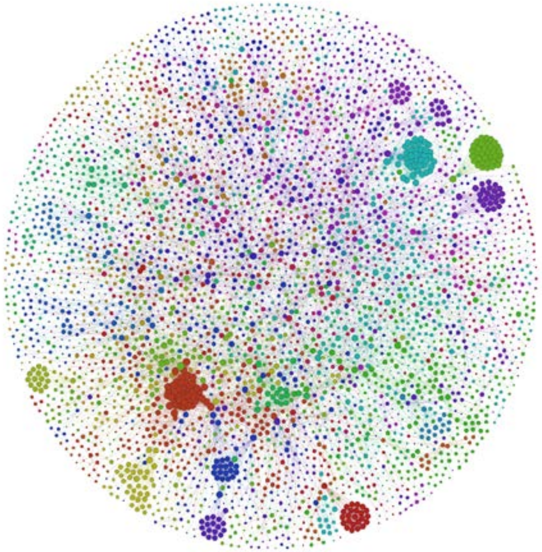
Identification of network communities

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

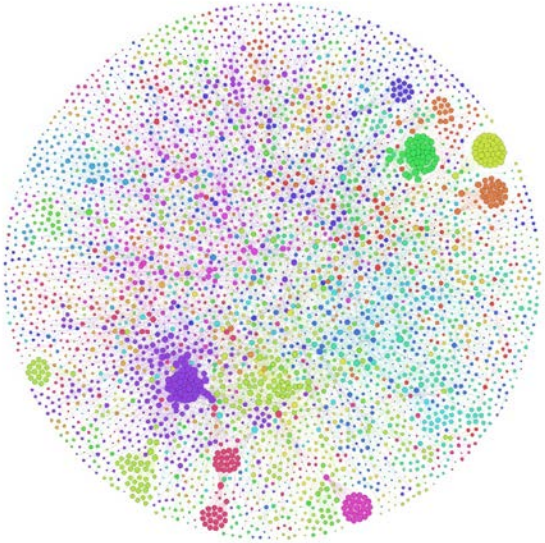
arXiv collaboration network (General Relativity): *N=4,158* nodes, 13,422 edges, *K = 36* [Leskovec'11]



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 $\mathcal{O}(d)$

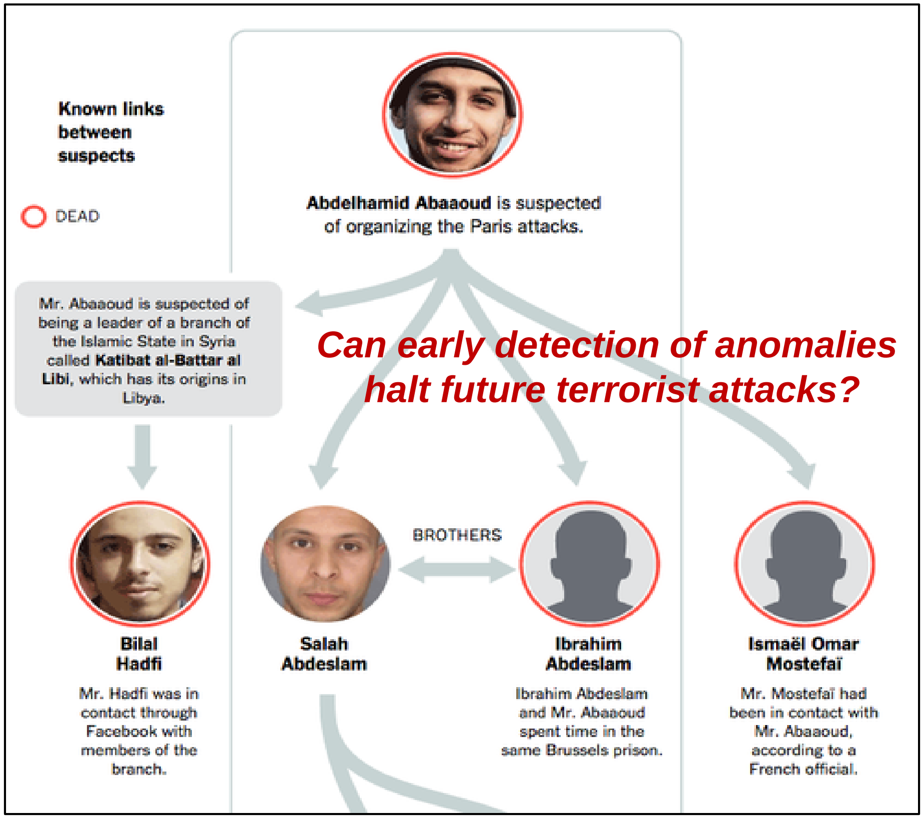
P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Spectral clustering of large-scale communities via random sketching and validation," *Proc. Conf. on Info. Science and Systems,* Baltimore, Maryland, March 18-20, 2015.

Roadmap

- ❑ Context and motivation
- ❑ Large-scale linear regressions
- ❑ Large-scale data and graph clustering
- ❑ Leveraging sparsity and low rank
  - ➢ Anomaly identification
  - ➢ Tensor subspace tracking
- ❑ Closing comments

Anomalies in social graphs

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



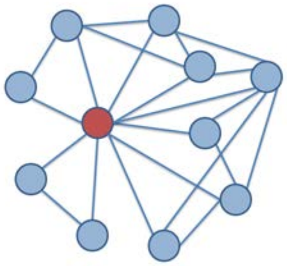
- **Examples**
  - E-mail spammers
  - Cybercriminals
  - Terrorist cells
- **Egonet features**
  - Degree, number of edges, centrality, betweeness, …

- **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

Egonet



- 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**

$$\mathbf{Y} = \mathbf{X} + \mathbf{O} + \mathbf{E}$$

Low-rank nominal features ($\mathbf{X}$) ***Sparse outlier matrix*** ($\mathbf{O}$)

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

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

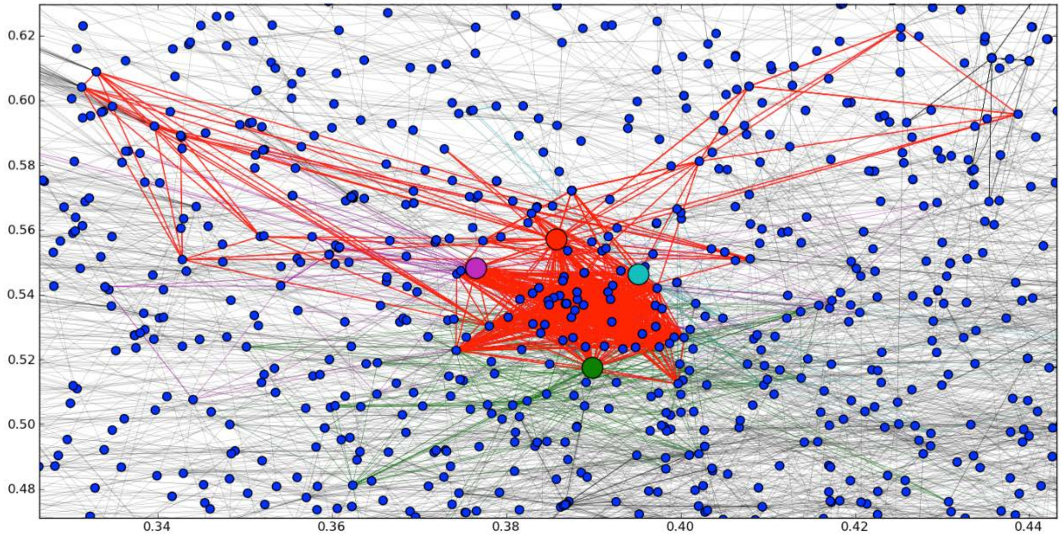
Robust low-rank component pursuit

- 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.

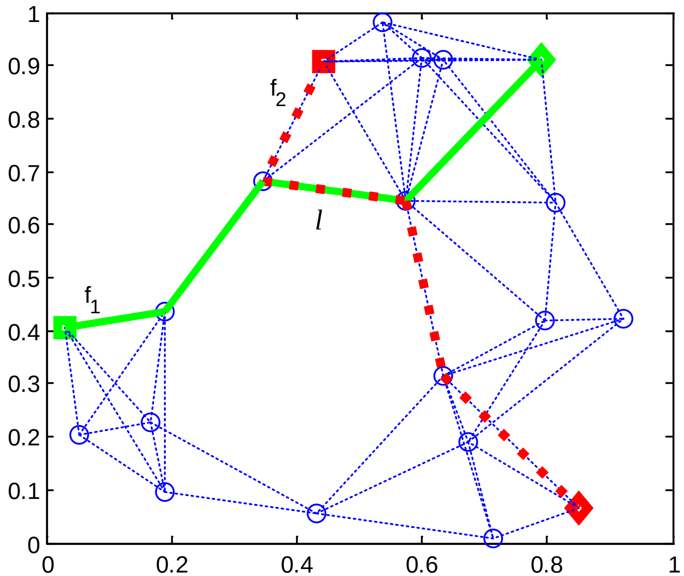
Modeling Internet traffic anomalies

- Anomalies: changes in origin-destination (OD) flows [Lakhina et al'04]
  - Failures, congestions, DoS attacks, intrusions, flooding
- Graph G (N, L) with N nodes, L links, and F flows ($F \gg L$); OD flow $z_{f,t}$
- Packet counts per link l and time slot t

Anomaly

$$y_{l,t} = \sum_{f=1}^{F} r_{l,f}(z_{f,t} + a_{f,t}) + v_{l,t}$$

$r_{l,f} \in \{0,1\}$

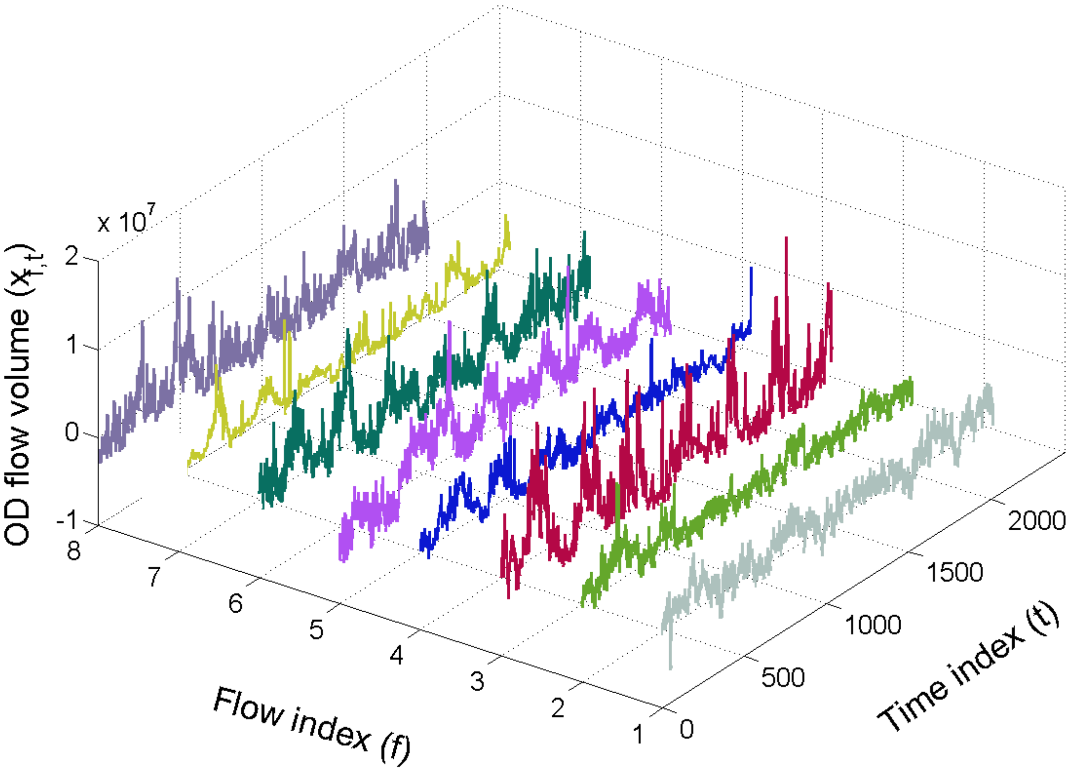


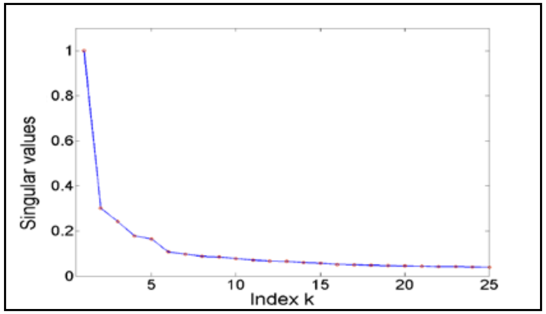
- Matrix model across T time slots: $\mathbf{Y} = \mathbf{R}(\mathbf{Z} + \mathbf{A}) + \mathbf{V}$

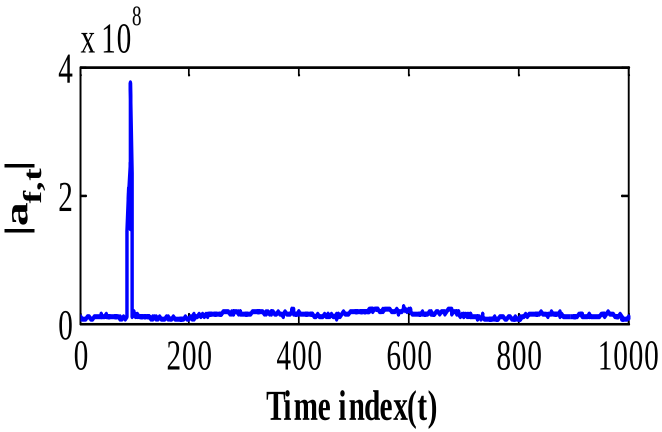
M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Transactions on Information Theory*, pp. 5186-5205, Aug. 2013.

Low-rank plus sparse matrices

- **Z** (and **X**:=**RZ**) low rank, e.g., [Zhang et al'05]; **A** is sparse across time and flows





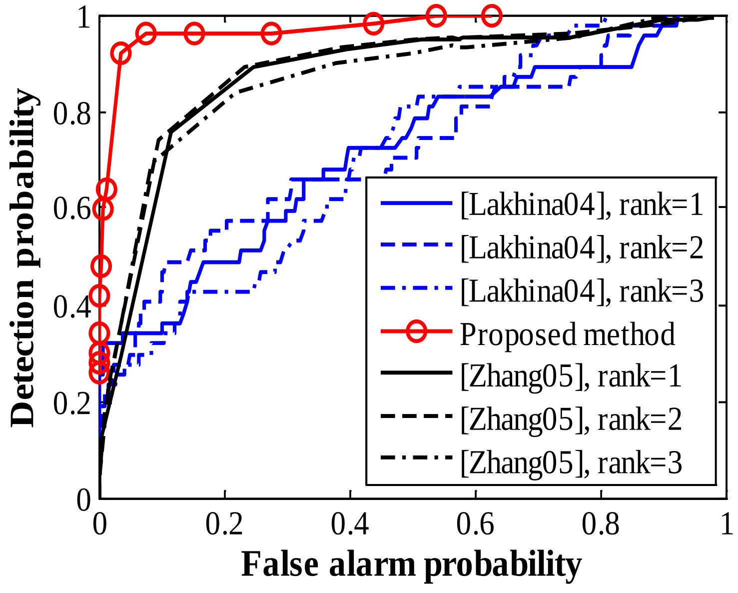


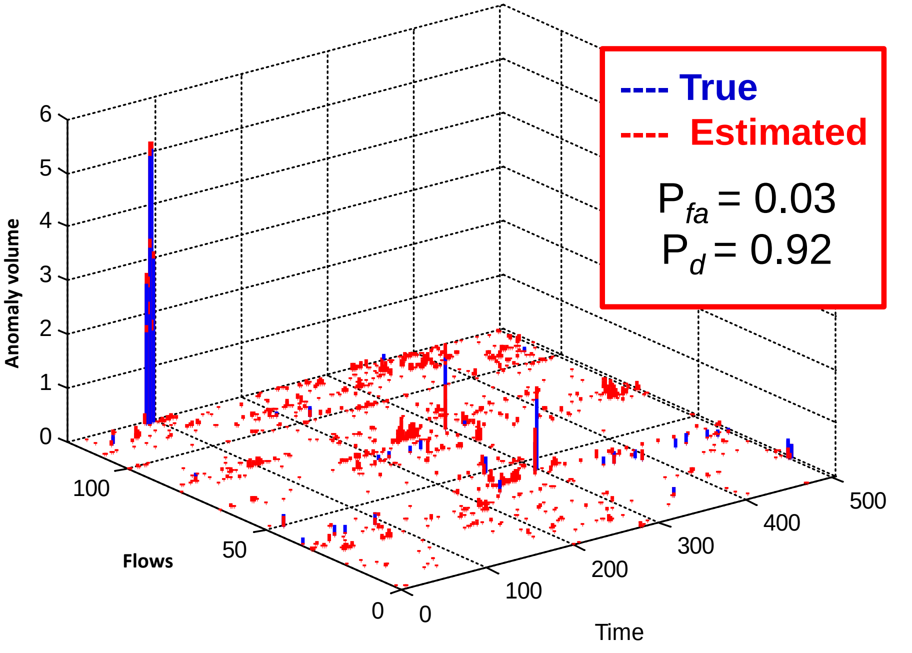
$$\{\hat{\mathbf{X}}, \hat{\mathbf{A}}\} = \arg\min_{\{\mathbf{X},\mathbf{A}\}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{A}\|_F^2 + \lambda_1\|\mathbf{A}\|_1 + \lambda_*\|\mathbf{X}\|_* \quad \text{(P1)}$$

Data: http://math.bu.edu/people/kolaczyk/datasets.html

Internet2 data

- Real network data, Dec. 8-28, 2003





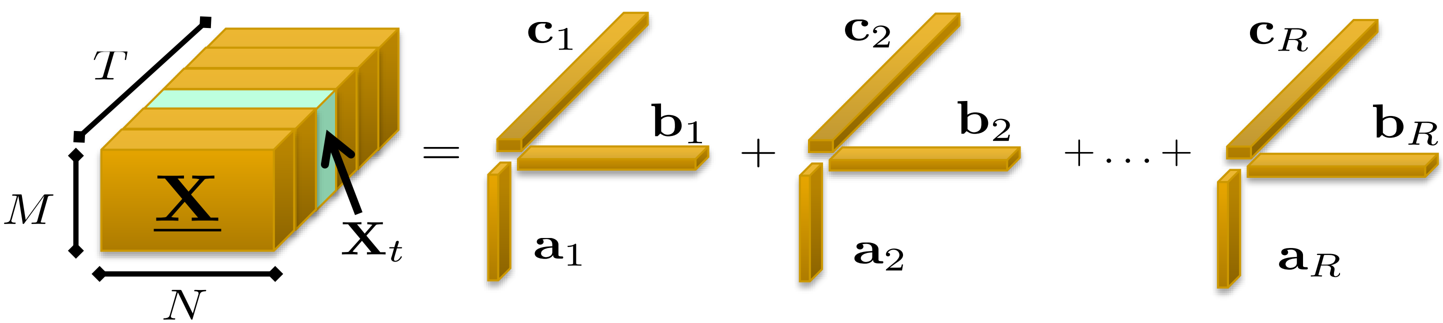
- Improved performance by leveraging sparsity and low rank
- Succinct depiction of the network health state across flows and time

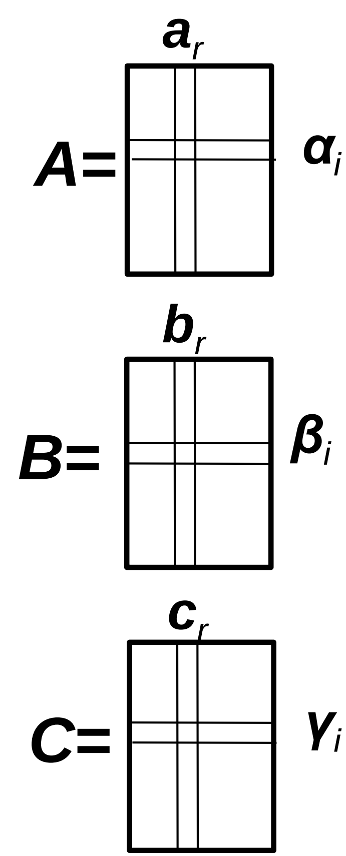
Data: http://www.cs.bu.edu/~crovella/links.html

From low-rank matrices to tensors

- Data cube $\underline{\mathbf{X}} \in \mathbb{R}^{M \times N \times T}$, e.g., sub-sampled MRI frames

$$\mathbf{Y}_t^{\Omega} \approx \mathcal{F}_{\Omega_t}(\mathbf{X}_t)$$





- PARAFAC decomposition per slab *t* [Harshman '70]

$$\mathbf{X}_t = \sum_{r=1}^{R} \gamma_{t,r} \mathbf{a}_r \mathbf{b}_r^{\top} = \mathbf{A}\mathrm{diag}(\boldsymbol{\gamma}_t)\mathbf{B}^{\top}$$

- Tensor subspace comprises *R* rank-one matrices $\{\mathbf{a}_r \mathbf{b}_r^{\top}\}_{r=1}^{R}$

Goal: Given streaming $\mathbf{Y}_t^{\Omega} \approx \mathcal{F}_{\Omega_t}(\mathbf{A}\mathrm{diag}(\boldsymbol{\gamma}_t)\mathbf{B}^{\top})$, learn the subspace matrices (***A***,***B***) recursively, and impute possible misses of ***Y***$_t$

J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Rank regularization and Bayesian inference for tensor completion and extrapolation," *IEEE Trans. on Signal Processing*, vol. 61, no. 22, pp. 5689-5703, Nov. 2013.

Online tensor subspace learning

- Image domain low tensor rank $\mathbf{Y}_t^{\Omega} \approx \mathcal{F}_{\Omega_t}(\mathbf{A}\mathrm{diag}(\boldsymbol{\gamma}_t)\mathbf{B}^{\top})$

$$(\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t) = \arg\min_{\mathbf{A},\mathbf{B}} \frac{1}{t}\sum_{\tau=1}^{t} \min_{\boldsymbol{\gamma}_\tau} \left\{ \|\mathbf{Y}_\tau^{\Omega} - \mathcal{F}_{\Omega_\tau}(\mathbf{A}\mathrm{diag}(\boldsymbol{\gamma}_\tau)\mathbf{B}^{\top})\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{\gamma}_\tau\|^2 \right\} + \frac{\lambda}{2t}(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$$

- Tikhonov regularization promotes low rank

Proposition [Bazerque-GG '13]: With $[\boldsymbol{\sigma}]_r = \|\mathbf{a}_r\|\|\mathbf{b}_r\|\|\mathbf{c}_r\|$

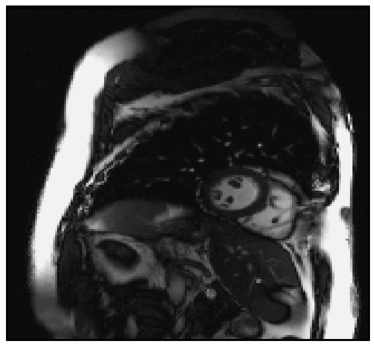
$$\|\boldsymbol{\sigma}(\underline{\mathbf{X}})\|_{2/3}^{2/3} = \min_{\{\mathbf{A}\mathbf{D}_t\mathbf{B}^T = \mathbf{X}_t\}} \left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\right)$$

- Stochastic alternating minimization; parallelizable across bases
- Real-time reconstruction (FFT per iteration) $\hat{\mathbf{X}}_t = \hat{\mathbf{A}}_t\mathrm{diag}(\hat{\gamma}_t)\hat{\mathbf{B}}_t^{\top}$

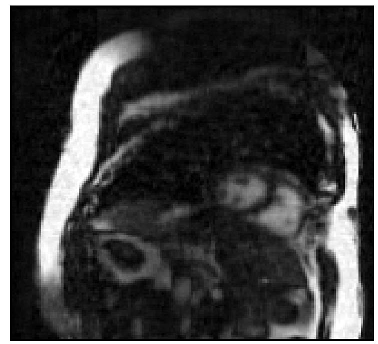
M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *IEEE Trans. on Signal Processing,* vol. 63, pp. 2663 - 2677, May 2015.

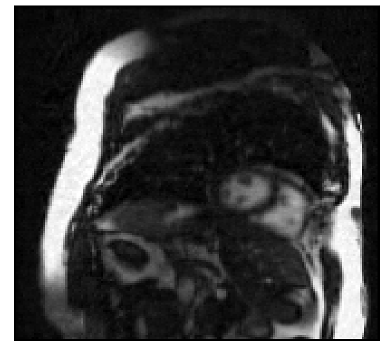
Dynamic cardiac MRI test

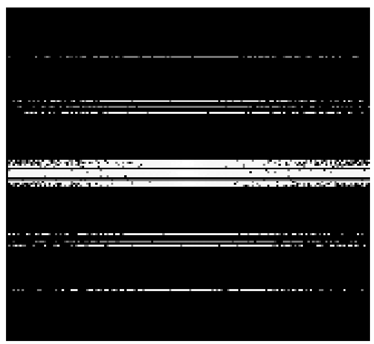
- *in vivo* dataset: 256 k-space 200x256 frames



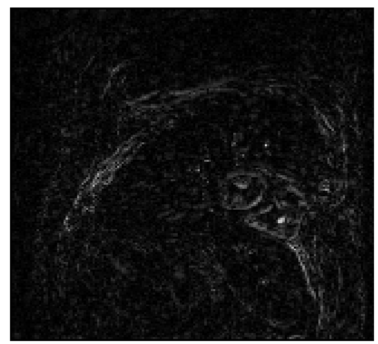
Ground-truth frame



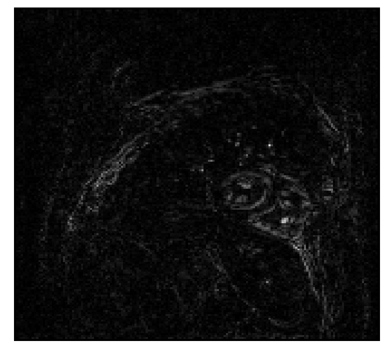




Sampling trajectory



R=100, 90% misses



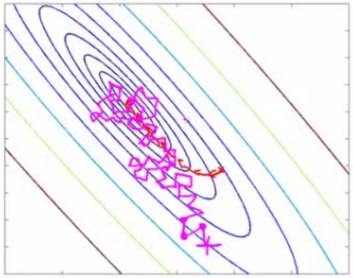
R=150, 75% misses

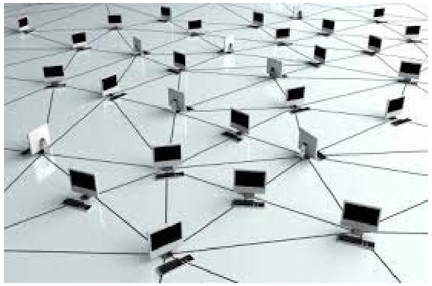
- Potential for accelerating MRI at high spatio-temporal resolution
- Low-rank $\mathcal{F}_{\Omega_t}(\mathbf{X}_t)$ plus $\mathcal{F}_{\Omega_t}(\mathbf{DS}_t)$ can also capture motion effects

M. Mardani and G. B. Giannakis, "Accelerating dynamic MRI via tensor subspace learning,“
Proc. of ISMRM 23rd Annual Meeting and Exhibition, Toronto, Canada, May 30 - June 5, 2015.

Closing comments

- **Large-scale learning**
  - Regression and tracking dynamic data
  - Nonlinear non-parametric function approximation
  - Clustering massive, high-dimensional data and graphs





- **Other key Big Data tasks**
  - Visualization, mining, privacy, and security
- **Enabling tools for Big Data**
  - Acquisition, processing, and storage
  - Fundamental theory, performance analysis
    decentralized, robust, and parallel algorithms
  - Scalable computing platforms
- **Big Data application domains …**
  - Sustainable Systems, Social, Health, and Bio-Systems, Life-enriching Multimedia, Secure Cyberspace, Business, and Marketing Systems …

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.