HPC and High-end Data Science for the Power Grid

Alex Pothen

August 3, 2018
Outline

1. High-end Data Science
2. Contingency Analysis
3. Data Anonymization
4. Parallel Oscillation Monitoring
PMUs in the U.S.

Phasor Measurement Units in the North American Grid as of March 2015 [1]

Challenges in High-end Data Analysis

- Computing is a major consumer of electricity, recent estimates about 10%.
- Data movement within a processor consumes more energy than arithmetic.
- Data movement from the instrument to the curation and archival sites, visualizations, researcher’s work spaces, etc. represents a significant use of energy.
- Hence energy is an overarching challenge for sustainable High-end Data Analysis (HDA).
- Solutions include:
  - Reduce computational costs by matching platforms to each stage of the scientific method
  - Reduce data movement costs by collocation, compression and caching
  - Reuse data through sharing, metadata, and catalogs
Challenges in High-end Data Analysis

- Data reduction is a fundamental pattern in the convergence of HPC and HDA.
  need new algorithms for loss-less and lossy compression of scientific data, and need algorithms that can work with the compressed data.
- New algorithms and software for HPC, data science, analytics, Machine learning, etc. are needed here.
- A major challenge is the convergence in the software ecosystem for HPC and HDA.
Challenges in High-end Data Analysis

- High volume data needs high performance computing
- Streaming data needs on-line algorithms
- Integration of heterogeneous data
  e.g., consumption forecasting in smart grids: smart meters, utility databases, weather data, microgrid sensor networks, etc.
- Anonymization of consumer data
Contingency Analysis

$N - x$ contingency analysis on power flow evaluates the stability of a power system by simulating the failures of $x$ out of $N$ transmission lines or generators (components).
Contingency Analysis

$N - x$ contingency analysis on power flow evaluates the stability of a power system by simulating the failures of $x$ out of $N$ transmission lines or generators (components).
Contingency Analysis

$N - x$ contingency analysis on power flow evaluates the stability of a power system by simulating the failures of $x$ out of $N$ transmission lines or generators (components).
Currently each $N - x$ contingency analysis is performed independently, but since the number of cases increases exponentially with $x$, this is computationally impractical. (For example, if $N = 3120$ and $x = 5$, there are in total $2.46 \times 10^{15}$ cases.)

Heuristics are needed for choosing the cases to be analyzed. To be able to analyze more cases, we have to speed up the solution time for each case.

We propose new algorithms for the problem by observing that only a principal submatrix of the system is changed when a component is removed upon failure.
Augmented Formulation

The augmented matrix system can be expressed as:

\[
\begin{bmatrix}
    A & AH & H \\
    H^T A & H^T AH & 0 \\
    H^T & 0 & 0
\end{bmatrix}
\begin{bmatrix}
    \hat{x}_1 \\
    \hat{x}_2 \\
    \hat{x}_3
\end{bmatrix}
= 
\begin{bmatrix}
    b \\
    H^T \hat{b}
\end{bmatrix}.
\]

- \(A\) is the original matrix;
- \(H\) is a submatrix of the identity matrix;
- \(\hat{A} = A - HEH^T\) updated matrix;
- \(m \ll n\);
- \(\hat{b}\) is the new right-hand-side of the updated rows.
Runnites

Power Grid Contingency Analysis of 777,646 Nodes

- PARDISO
- LUSOL
- CHOLMOD
- augmented (Direct)
- augmented (GMRES)
## Adaptive Anonymity

<table>
<thead>
<tr>
<th>Users</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_1$</td>
</tr>
<tr>
<td>$U1$</td>
<td>1</td>
</tr>
<tr>
<td>$U2$</td>
<td>1</td>
</tr>
<tr>
<td>$U3$</td>
<td>0</td>
</tr>
<tr>
<td>$U4$</td>
<td>0</td>
</tr>
<tr>
<td>$U5$</td>
<td>1</td>
</tr>
<tr>
<td>$U6$</td>
<td>1</td>
</tr>
</tbody>
</table>
Adaptive Anonymity

<table>
<thead>
<tr>
<th>Users</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_1$</td>
</tr>
<tr>
<td>$U1$</td>
<td>1</td>
</tr>
<tr>
<td>$U2$</td>
<td>1</td>
</tr>
<tr>
<td>$U3$</td>
<td>0</td>
</tr>
<tr>
<td>$U4$</td>
<td>0</td>
</tr>
<tr>
<td>$U5$</td>
<td>1</td>
</tr>
<tr>
<td>$U6$</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Users</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_1$</td>
</tr>
<tr>
<td>$U1$</td>
<td>1</td>
</tr>
<tr>
<td>$U2$</td>
<td>1</td>
</tr>
<tr>
<td>$U3$</td>
<td>0</td>
</tr>
<tr>
<td>$U4$</td>
<td>0</td>
</tr>
<tr>
<td>$U5$</td>
<td>1</td>
</tr>
<tr>
<td>$U6$</td>
<td>1</td>
</tr>
</tbody>
</table>
Adaptive Anonymity with Approximation Algorithms

<table>
<thead>
<tr>
<th>Prob.</th>
<th>Inst.</th>
<th>Feat.</th>
<th>b−EdgeCover</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time Util.</td>
</tr>
<tr>
<td>Census1990</td>
<td>158K</td>
<td>68</td>
<td>9m 90%</td>
</tr>
<tr>
<td>PokerHands</td>
<td>500K</td>
<td>95</td>
<td>2h 17m 84%</td>
</tr>
<tr>
<td>CMS</td>
<td>1M</td>
<td>512</td>
<td>10h 33m 81%</td>
</tr>
</tbody>
</table>

Earlier Algorithms

Belief propagation: used exact $b$-Matching algorithm, could solve problems with few thousand instances only.

The approximate edge cover approach is two to three orders of magnitude faster on those smaller problems.
Adaptive Anonymity Computation on Cori

![Graph showing strong scaling results for Cori@NERSC with distributed memory. The graph plots runtime (Sec) against the number of cores (128 to 8192). Lines represent different datasets: UCI_Adult, USCensus1990, Poker_hands, CMS17, and Ideal. The lines show decreasing runtime with increasing number of cores, indicating efficient scaling.]
Outline

1. High-end Data Science
2. Contingency Analysis
3. Data Anonymization
4. Parallel Oscillation Monitoring
PMUs in the U.S.

Stochastic Subspace Identification – Covariance

\[
\begin{align*}
\begin{cases}
x_{k+1} = Ax_k + w_k \\
y_k = Cx_k + v_k
\end{cases},
\end{align*}
\]

\[
y_p = \begin{bmatrix}
y_0 & y_1 & \cdots & y_{2-1} \\
y_1 & y_2 & \cdots & y_2 \\
\vdots & \vdots & \ddots & \vdots \\
y_{2-1} & y_2 & \cdots & y_{2-1}
\end{bmatrix}, \quad y_f = \begin{bmatrix}
y_1 & y_{1+1} & \cdots & y_{1+2-1} \\
y_{1+1} & y_{1+2} & \cdots & y_{1+2} \\
\vdots & \vdots & \ddots & \vdots \\
y_{1+2-1} & y_{1+2} & \cdots & y_{1+2-1}
\end{bmatrix},
\]

\[
H = Y_f Y_p^T = O_t G,
\]

\[
O_t = \begin{bmatrix}
C & CA & CA^2 & \cdots & CA^{l-1}
\end{bmatrix}^T, \quad G \equiv E \{ x_k y_k^T \},
\]

\[
H = O_t G = USV^T,
\]

\[
O_t = U_1 \times S_1^{1/2},
\]

\[
A = Q_t^\dagger \overline{O}_t, \quad C = O_t (1: l),
\]

\[
A_c = \frac{1}{T_s} \log A, \quad C_c = C,
\]

Mode Freq. and Damping: eigenvalues of \( \hat{A}_c \),

ModeShape: \( \phi_i = \hat{C}_c \phi_i^m \).
### Case 1 – 102 Voltage Phase Angles

#### Computational Time Comparison

<table>
<thead>
<tr>
<th>Mac. No.</th>
<th>Full SVD (sec)</th>
<th>Randomized SVD ( h = 20, q = 1 )</th>
<th>Lanczos SVD ( k = 16 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Strat. 1 sec (speedup)</td>
<td>Strat. 2 sec (speedup)</td>
</tr>
<tr>
<td>1</td>
<td>112.075</td>
<td>105.826 (1.06x)</td>
<td>206.8 (542x)</td>
</tr>
<tr>
<td>2</td>
<td>125.807</td>
<td>121.007 (1.04x)</td>
<td>237.9 (529x)</td>
</tr>
<tr>
<td>3</td>
<td>138.236</td>
<td>135.304 (1.02x)</td>
<td>298.4 (463x)</td>
</tr>
<tr>
<td>4</td>
<td>170.518</td>
<td>168.617 (1.01x)</td>
<td>304.4 (560x)</td>
</tr>
</tbody>
</table>
HDA and HPC

Data Anonymization

Data Reduction for Oscillation Monitoring
Contingency Analysis