Functional Singular Value Decomposition

Lingsong Zhang
October 6, 2005
Email: lszhang@email.unc.edu

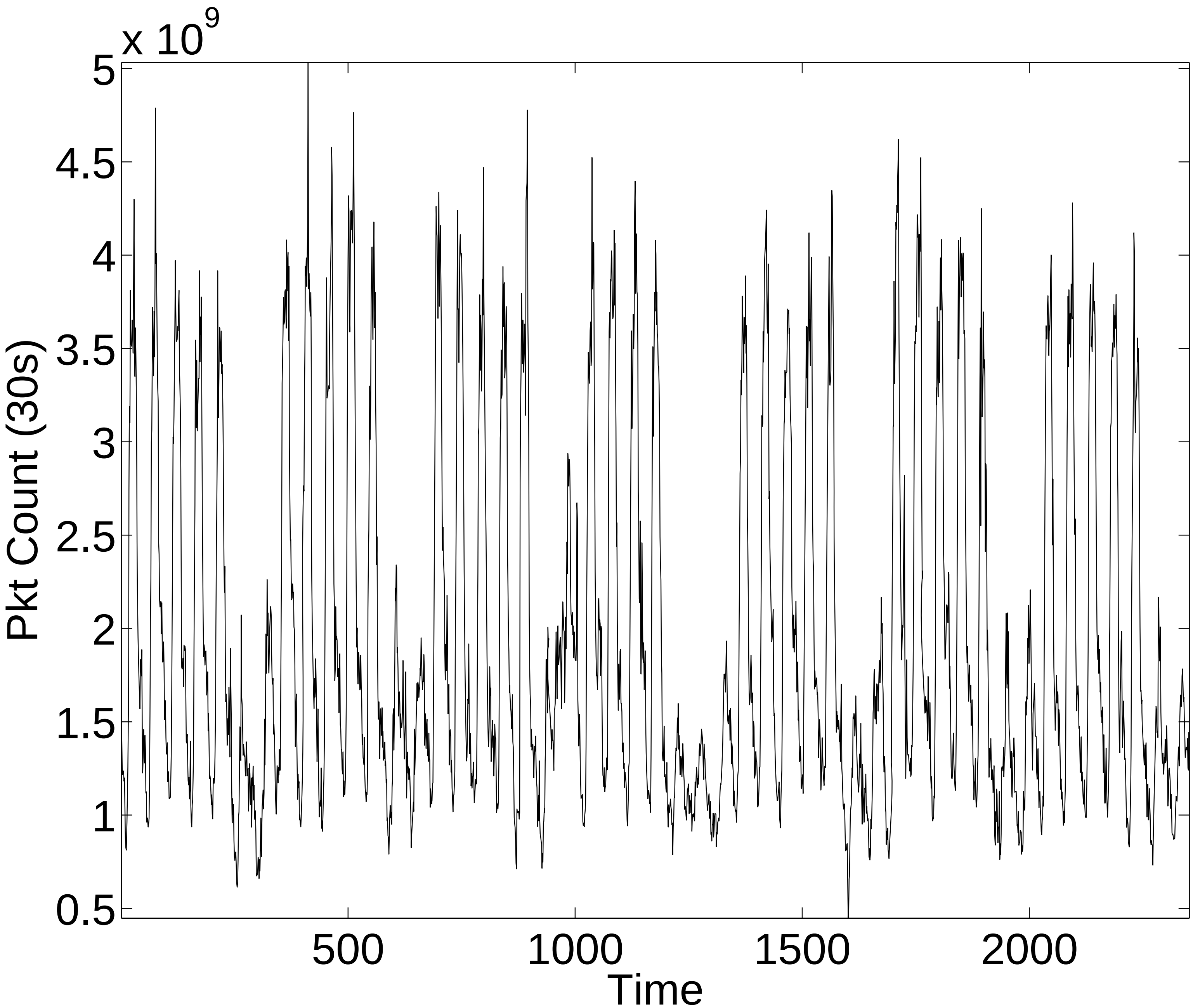
Advised by
J. S. Marron, Zhengyuan Zhu and Haipeng Shen

Outline

- Motivating example - Network traffic data
- Visualization methods
- SVD and PCA (If time permits)
- Future work

Motivating example

- Internet traffic data
- UNC campus, main Internet link of campus to outside
- Packet counts data
- Half-an-hour bin size
- 49 days, covered fully 7 weeks
- June 9, 2003 – July 27, 2003
- Cover two summer sessions of UNC summer school



Time series plot of the 49 days packet count data, bin size half an hour

Main Observations

- 49 spikes, clear daily pattern
- Weekly pattern
- Weekday-weekend effect

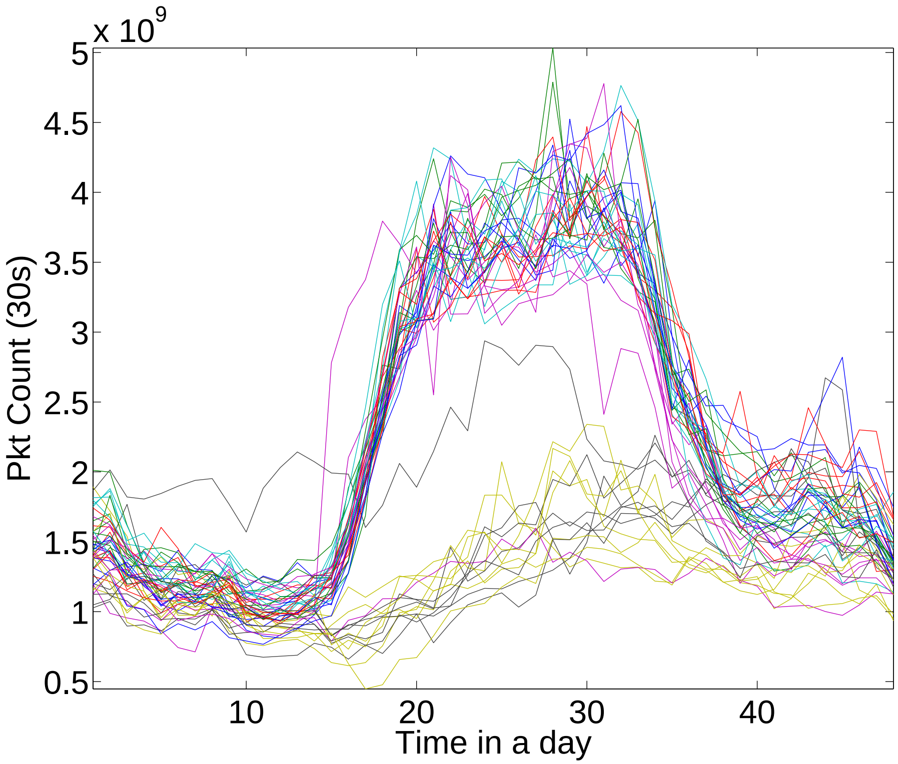
Matrix view of the data

- Rearrange the data as a 49×48 matrix
- Days in rows, Time-of-day in columns

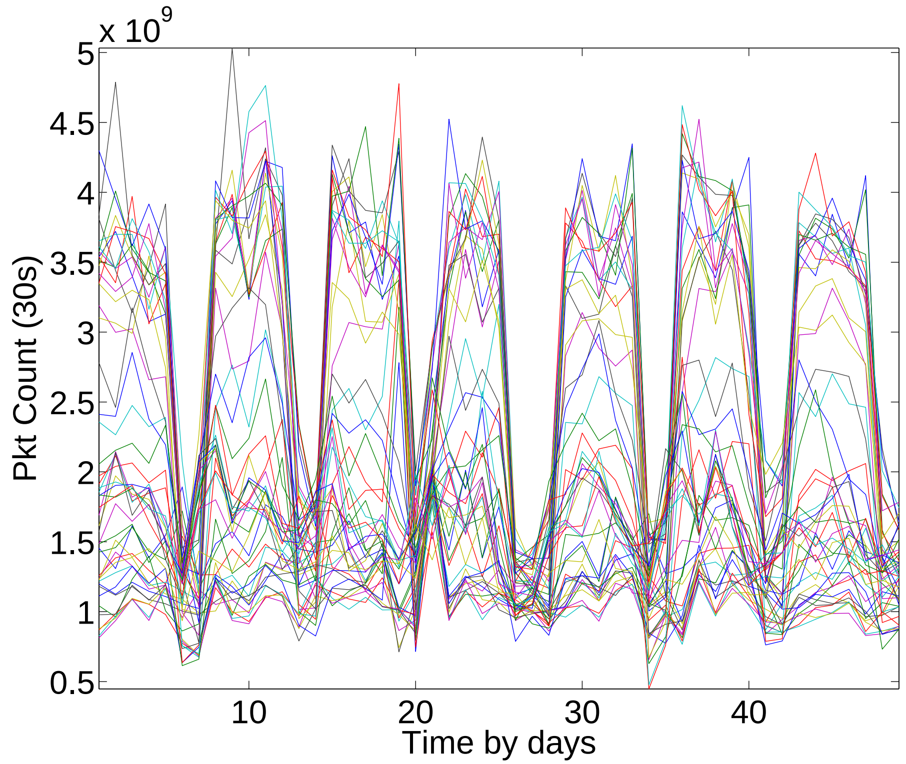
i.e. Rearrange the (x_1, x_2, ⋯, x_2352) into

$$\begin{array}{cc} & \begin{array}{cccc} \phantom{x_{2305}} & \text{time-of-day} & & \phantom{x_{2352}} \end{array} \\ \text{day} & \begin{pmatrix} x_1 & x_2 & \cdots & x_{48} \\ x_{49} & x_{50} & \cdots & x_{96} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2305} & x_{2306} & \cdots & x_{2352} \end{pmatrix} \end{array}$$

Two different FDA views



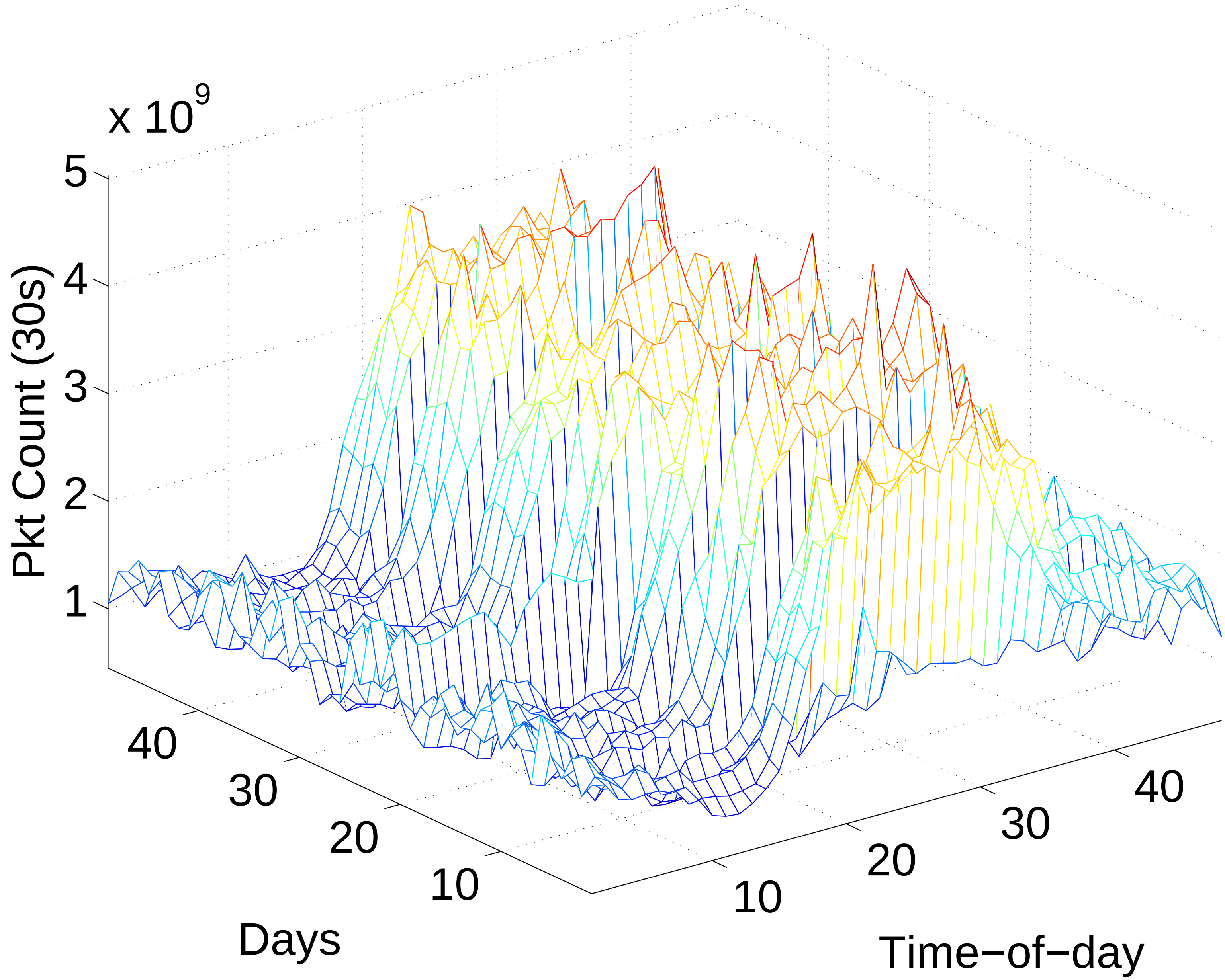
(a) Treat daily shapes as (functions) curves



(b) Treat cross-day time series as (functions) curves

Motivation of matrix view

- Show the daily shapes and Cross day time series at the same time.
- Combining both Functional Data Analysis views



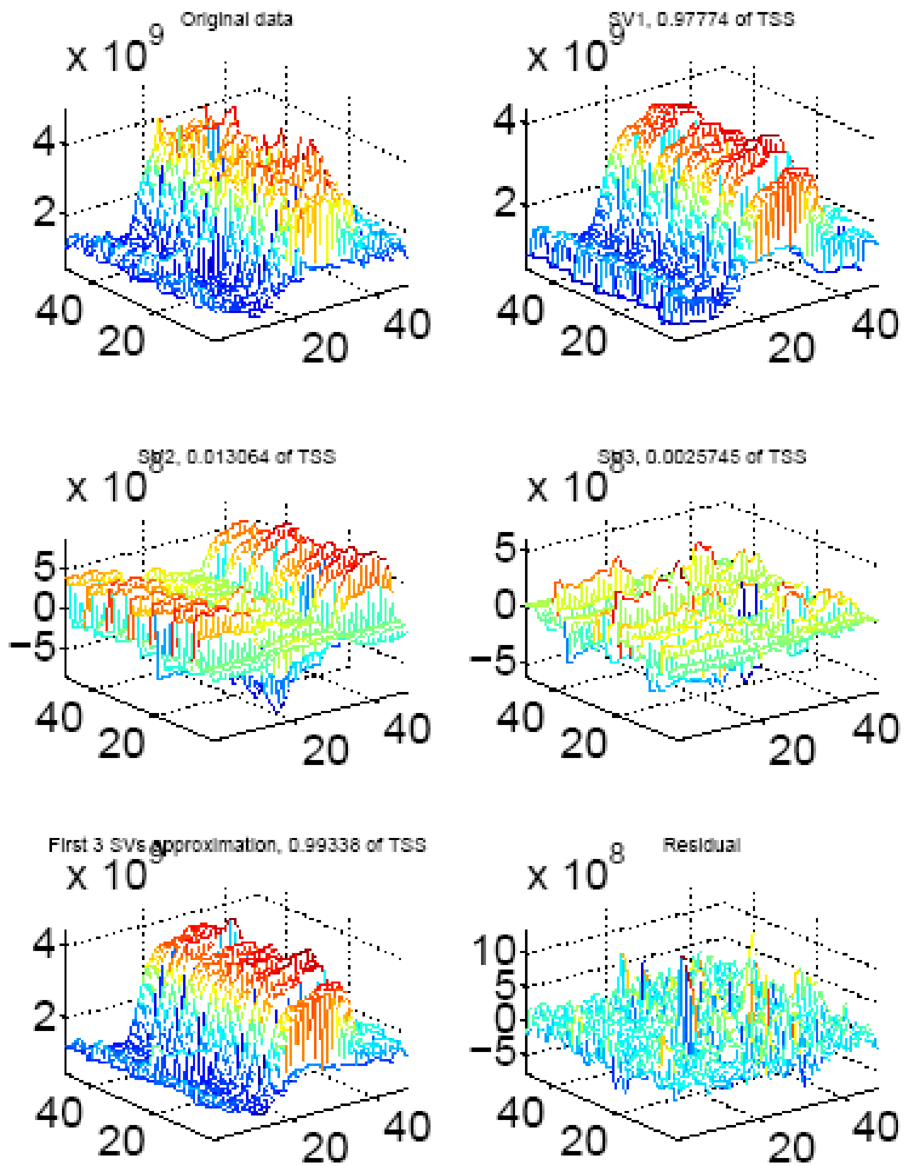
Mesh plot of the half hour network traffic data (matrix)

Decomposition into Modes of Variation

- PCA is a typical FDA method, SVD is very similar

- SVD can be done directly to the data matrix, might help to explore the original data matrix.

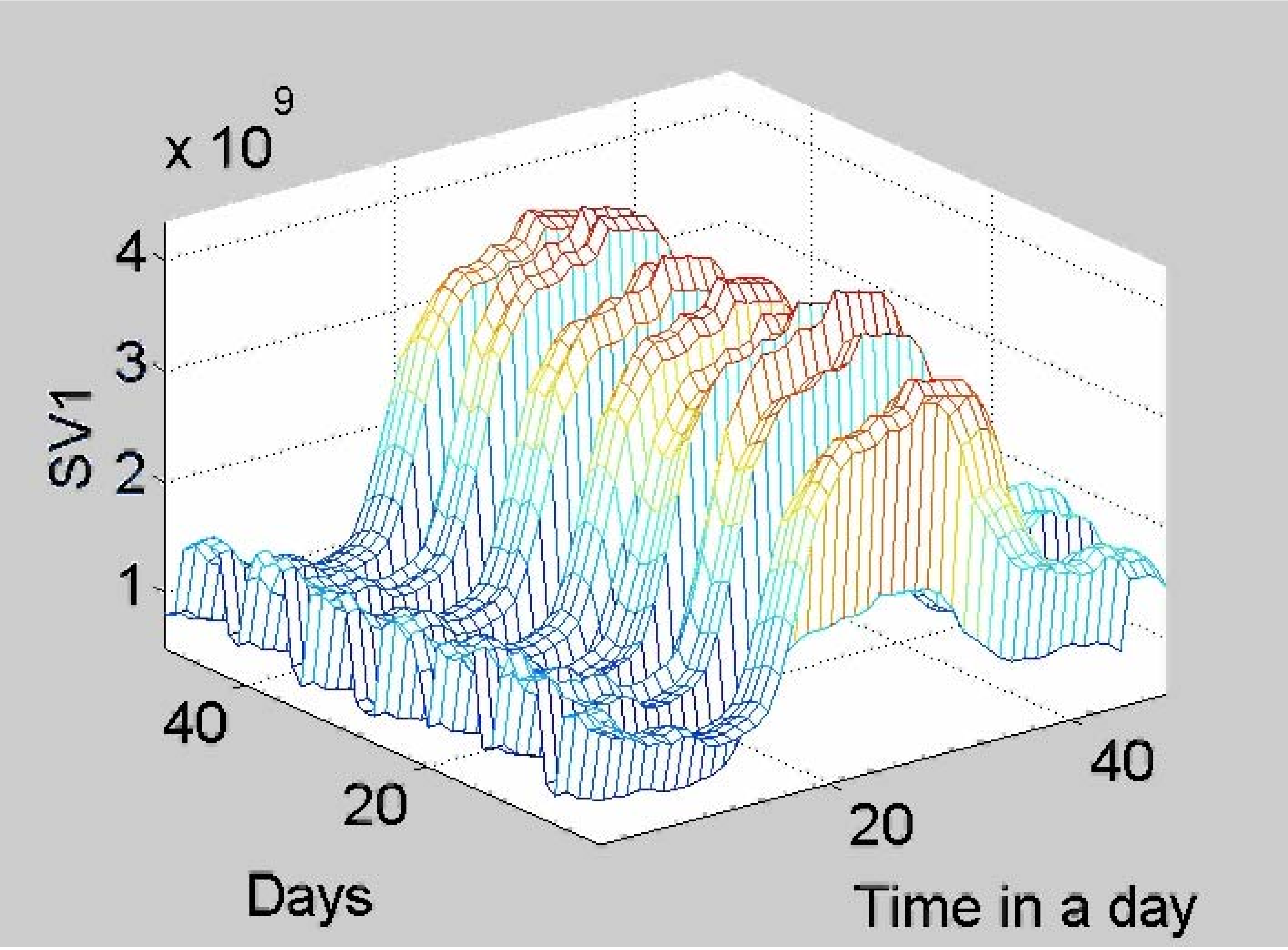
- (Surface plots of network traffic data)


Original data
SV1, 0.97774 of TSS
SV2, 0.013064 of TSS
SV3, 0.0025745 of TSS
First 3 SVs approximation, 0.99336 of TSS
Residual

Major Modes of Variations

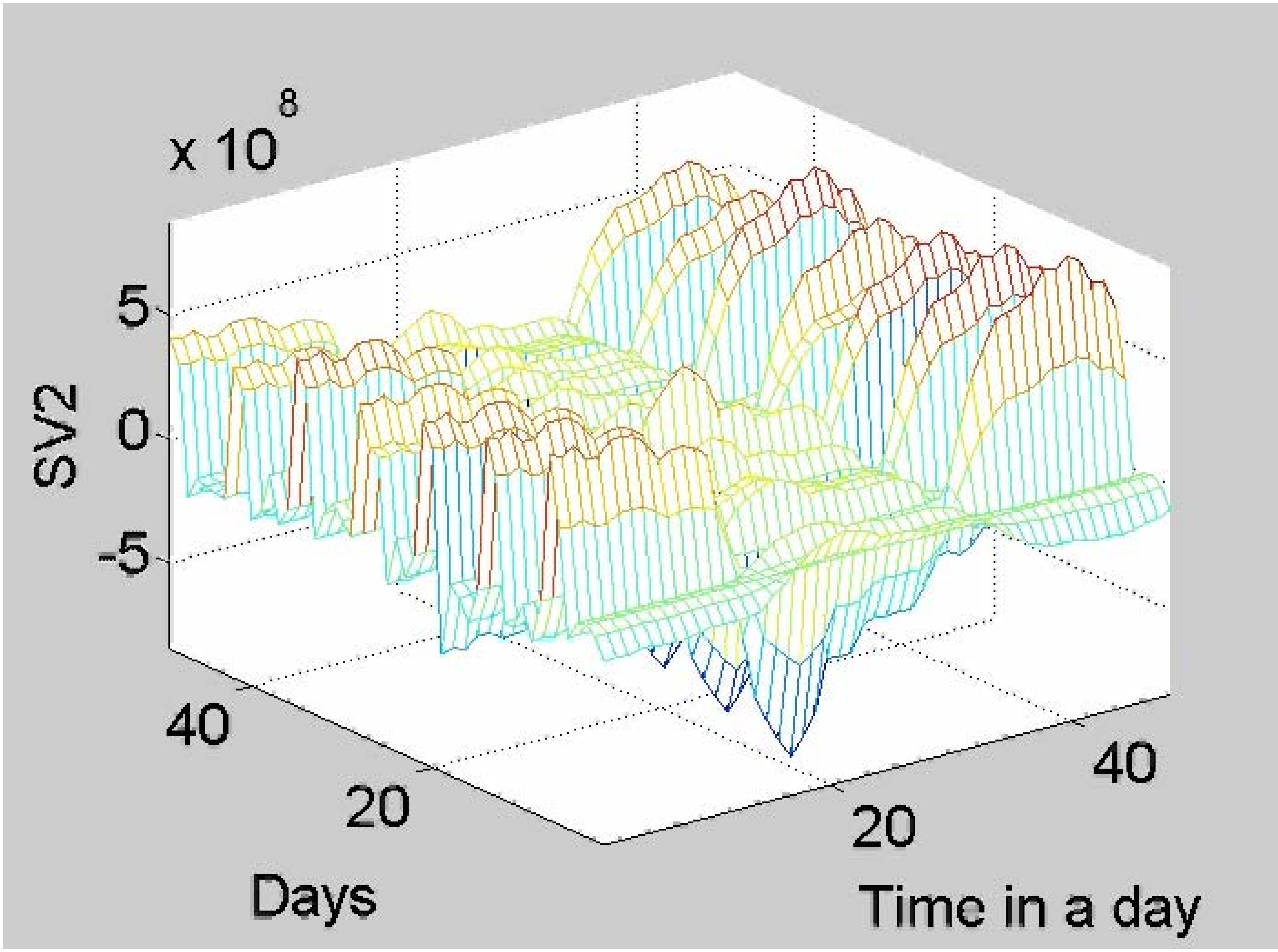
- First Component
  - Smoothed version of original data
  - Daily shapes
  - Weekly pattern
- Second Component
  - Weekday-Weekend effect
  - Weekday and Weekend might not share the same shapes
- Third Component
  - Outliers
- Residual
  - Seems to be noise

Different angles might help



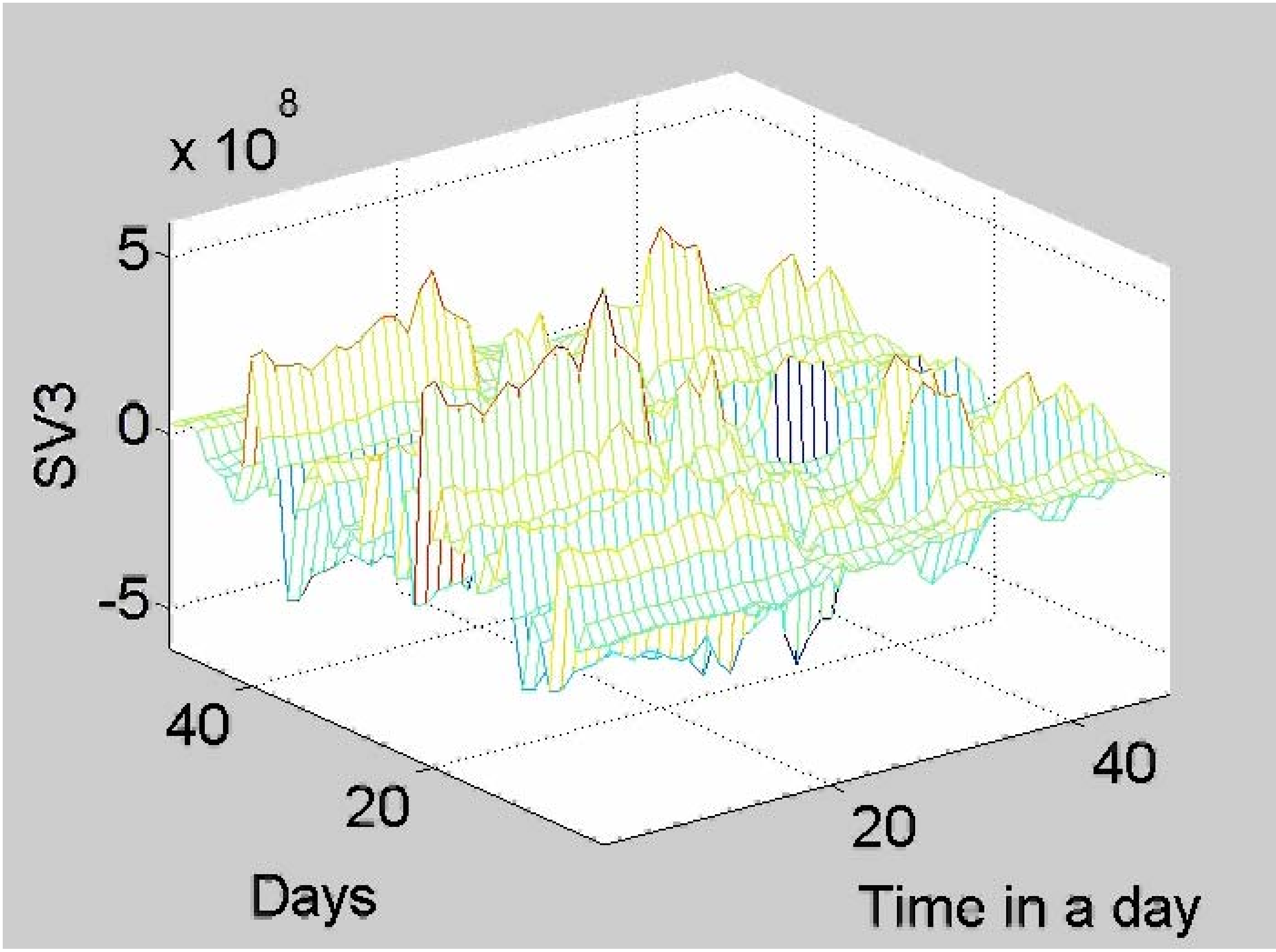
SVD Rotation Movie for SV1

Different angles might help



SVD Rotation Movie for SV2

Different angles might help



SVD Rotation Movie for SV3

Rotation Movies for network data

- First component
  - Common daily shapes, clearly weekly pattern
  - Long Weekend, July 4
- Second component
  - Weekday-weekend effect
  - July 27
- Third component
  - Outlier effect

Singular Value Decomposition

- Decompose the data matrix into several rank 1 (matrix) components.
- Each component has both column and row features.
- Surface plots highlight those features simultaneously.

Singular Value Decomposition

$$
\begin{aligned}
X &= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}_{m \times n} \\
&= USV^T \\
&= \begin{pmatrix} u_{11} & \boxed{u_{12}} & \cdots & u_{1r} \\ u_{21} & u_{22} & \cdots & u_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mr} \end{pmatrix}_{m \times r} \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_r \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1r} \\ v_{21} & v_{22} & \cdots & v_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nr} \end{pmatrix}_{n \times r}^T \\
&= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \end{pmatrix} \mathbf{S} \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \end{pmatrix}^T \\
&= s_1 \mathbf{u}_1 \mathbf{v}_1^T + \boxed{s_2 \mathbf{u}_2 \mathbf{v}_2^T} + \cdots + s_r \mathbf{u}_r \mathbf{v}_r^T
\end{aligned}
$$

SVD Component

Singular Column

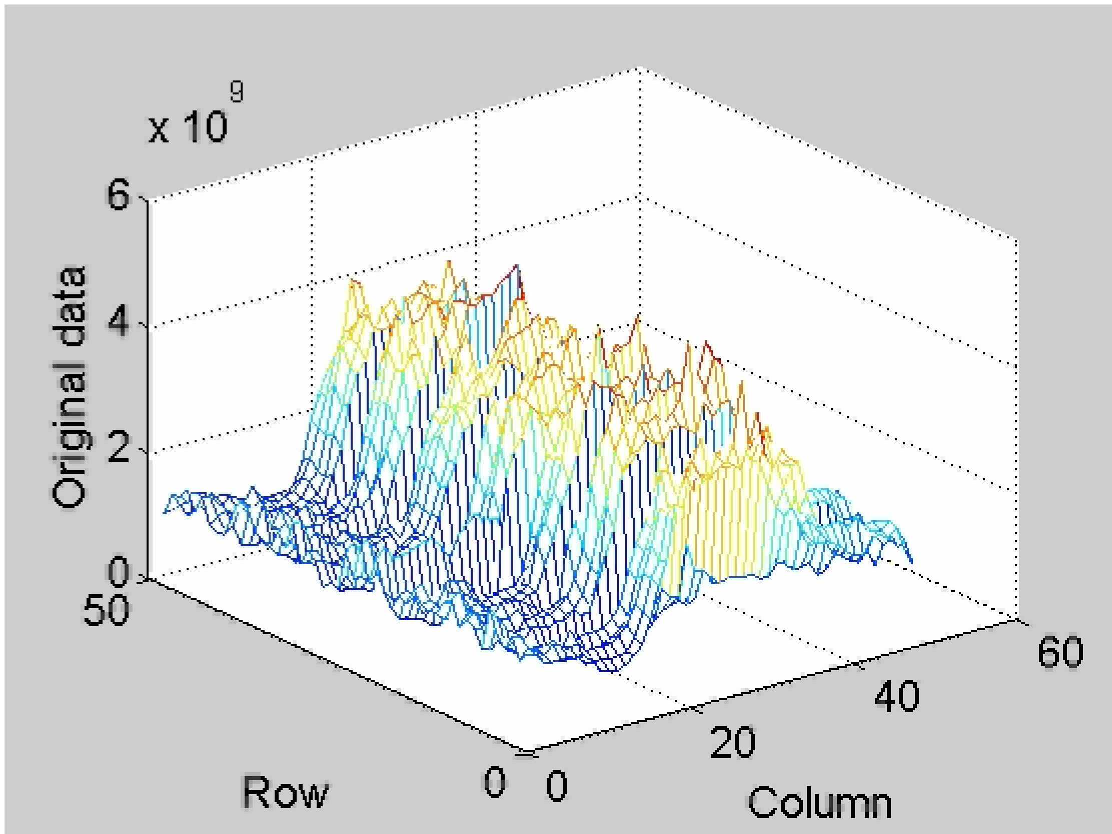
Singular Value

Singular Row

Singular Value Decomposition

- Let $\{r_i\}$, $\{c_j\}$ be the row and column vectors of the matrix X respectively
  - Singular Columns $\{u_i\}$ form an orthonormal basis for the column vector space
  - Singular Rows $\{v_i\}$ form an orthonormal basic for the row vector space
- The first k ($K \leq r$ =rank(X)) SVD components provide the best rank k approximation of the data matrix X

SVD curve movie



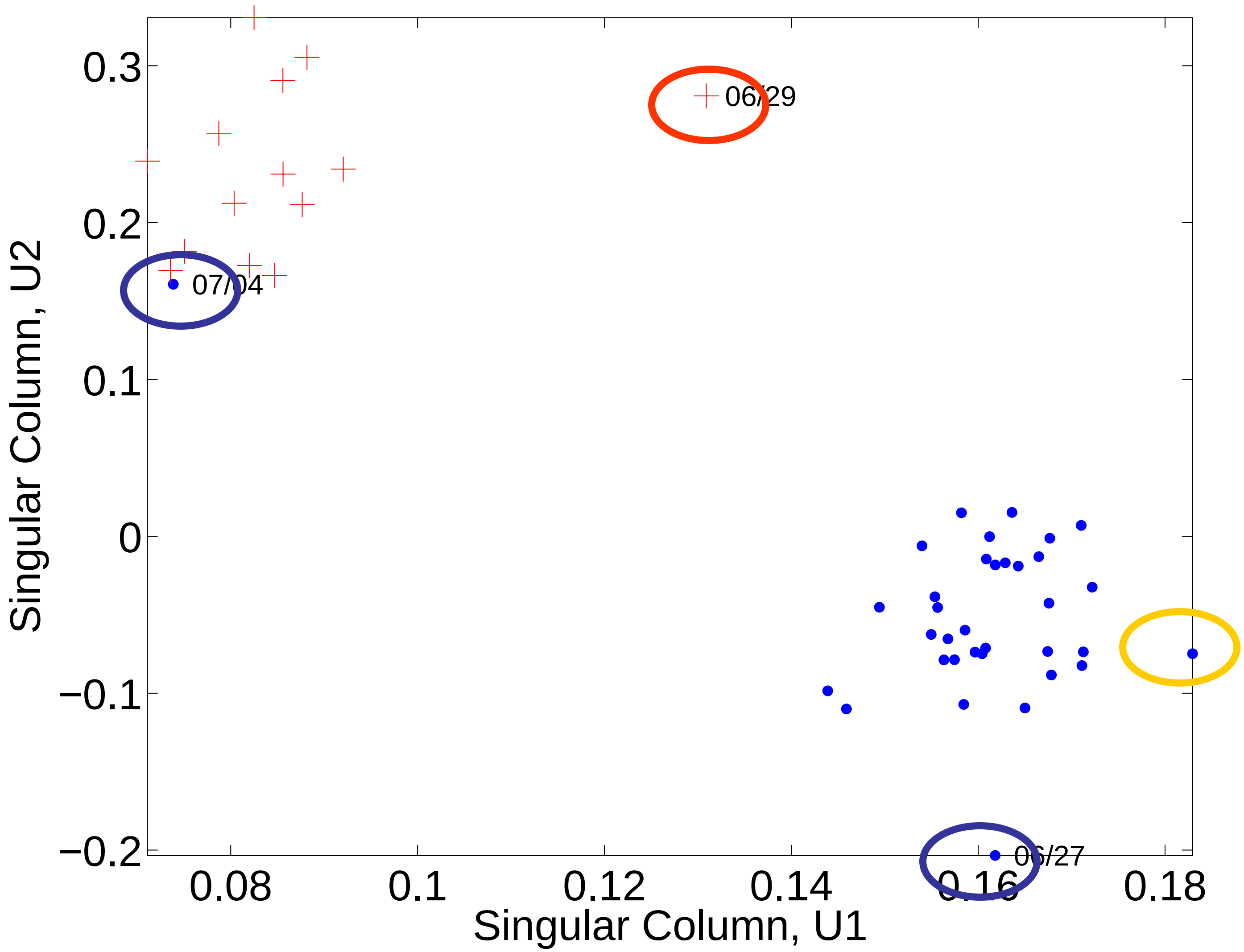
SVD curve movie for the network traffic data

SVD curve movie

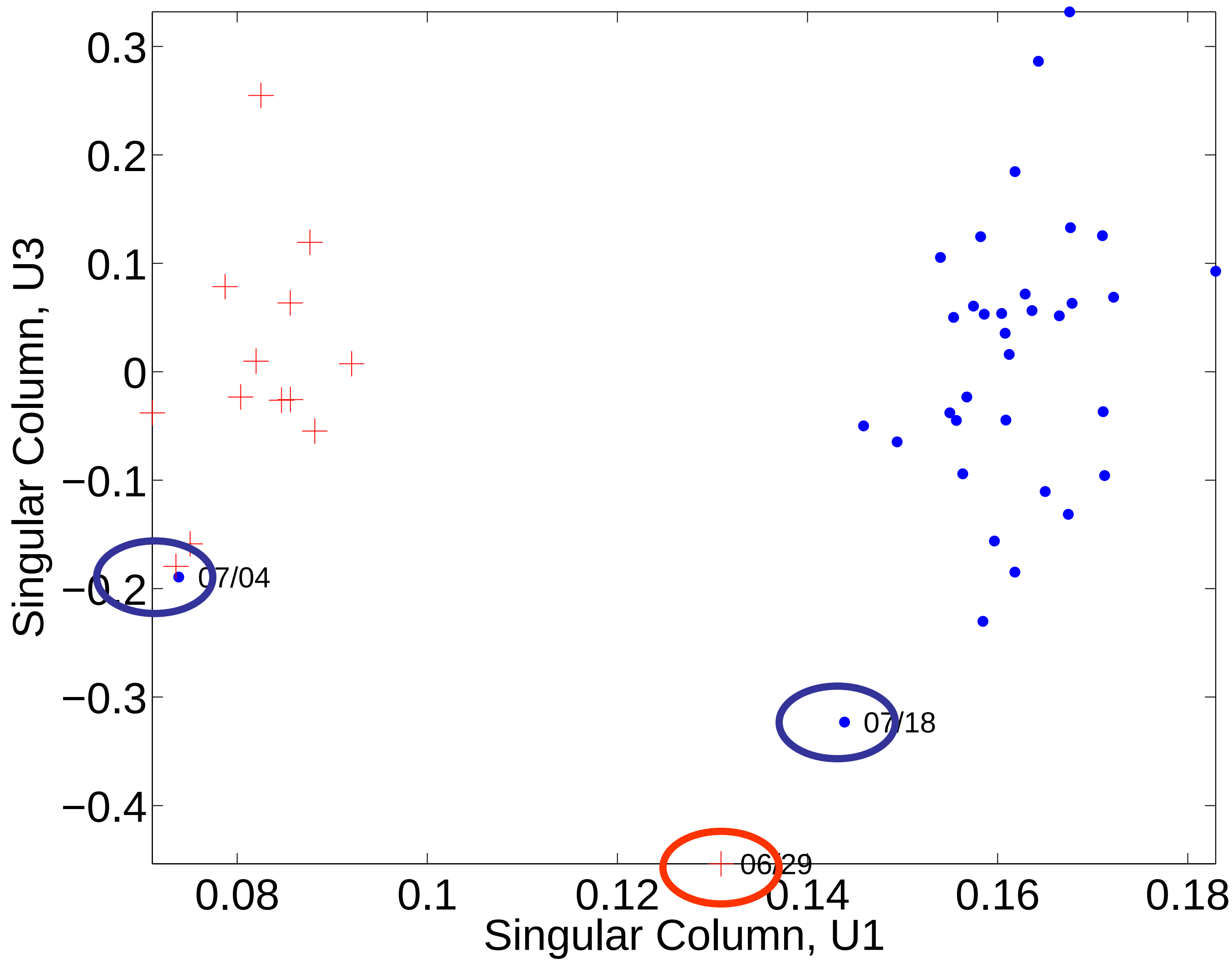
- Help to understand what SVD component is from
  - Outer product of singular column and singular row
- Show time varying features
- SVD curve movie for the third component
  - June 29, First Sunday of the Second Session
  - June 27, Last registration day for the Second Session
  - July 18, With 8 minutes missing data gap

Other Visualization Methods

- Scatter plots of singular columns
  - Treat the daily shapes as the functional curves, it is like the projection to the subspace spanned by the PCs.
  - Will help to find some special days.



Scatter plot of singular columns u_1 vs. u_2



Scatter plot of singular columns u_1 vs. u_3

Matlab software is available

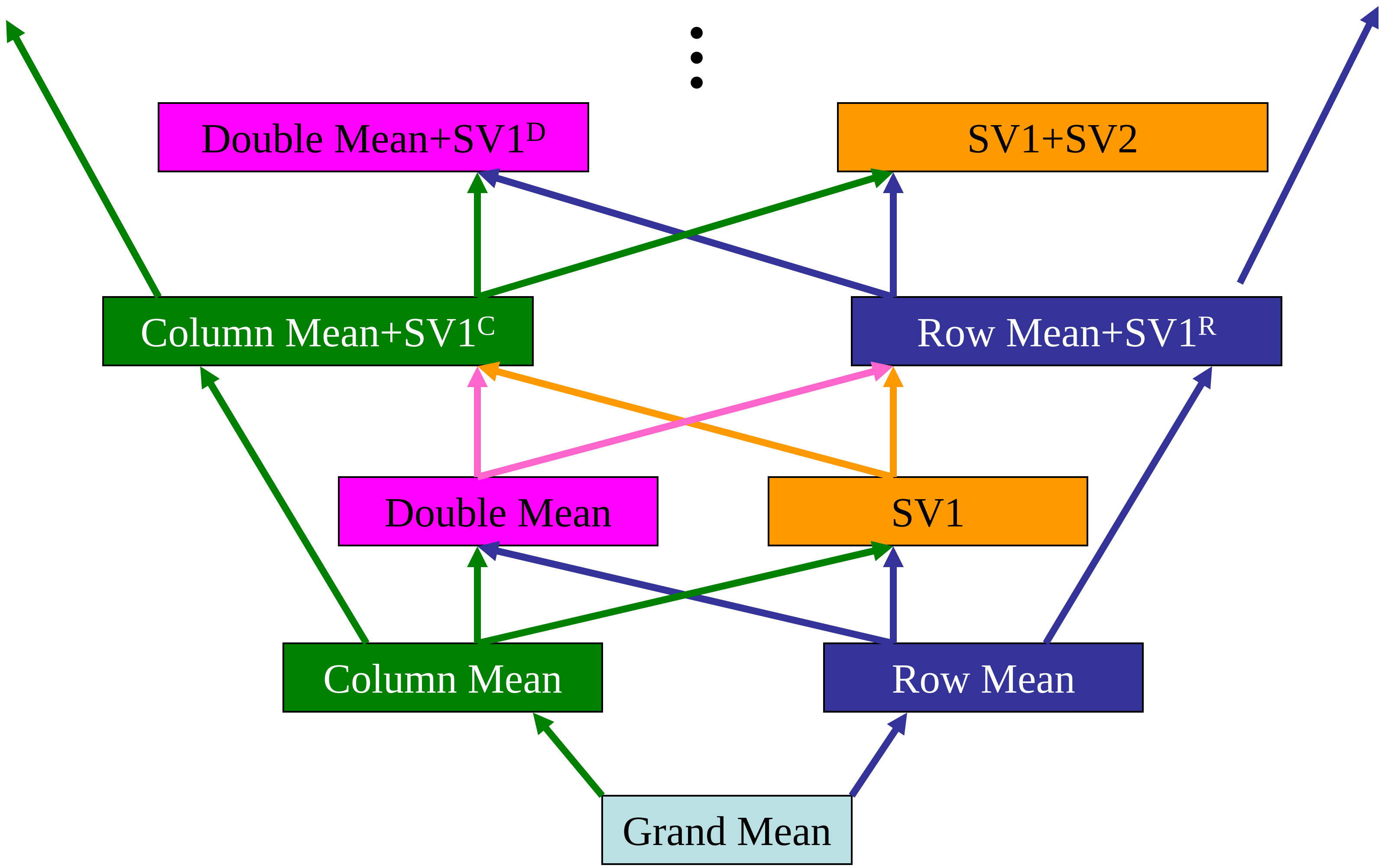
http://www.unc.edu/~lszhang/research/network/SVDmovie/

- SVD surface plots
- SVD rotation movie
- SVD curve movie
- Zoomed version of SVD curve movie
- Some plots and movies for the network traffic data and a chemometrics data

PCA and SVD

- Connections
  - If X is column centered at 0 (i.e. Column means are zeros), PCA is the factorization of X^TX.
  - SVD helps to get the PCs.
- Differences ?
  - Different factorization
    - PCA is the factorization of X^TX (covariance matrix)
    - SVD is the factorization of X (original data matrix)
    - Dual PCA is the factorization of XX^T
  - Recentering?
    - Why column centered at Zero?
    - Four types of centering: None, Column, Row and Both?

Approximation View



Four types of recentering

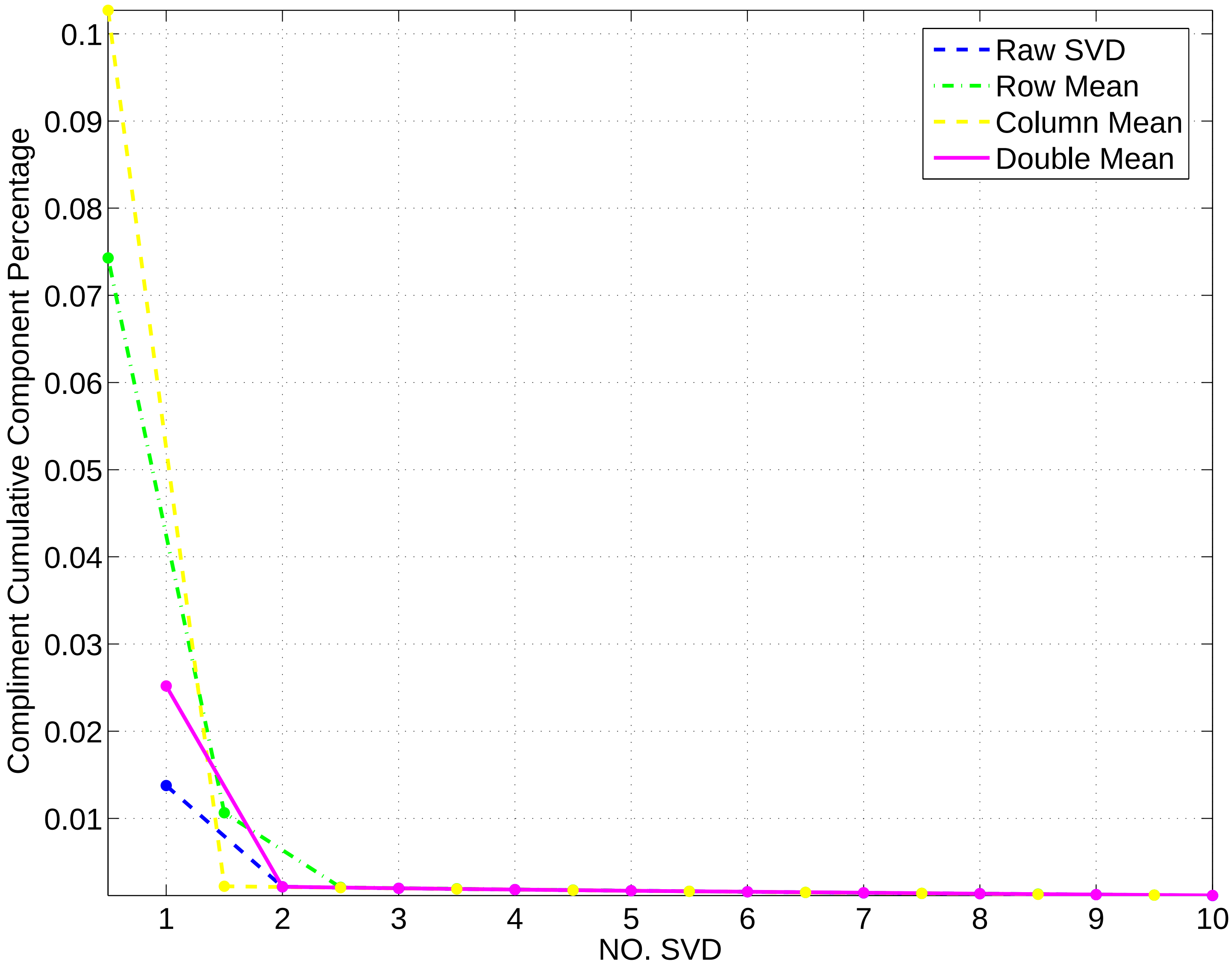
- SVD with no recentering is the best rank *k* approximation

- SVD with column recentering or row recentering are sub-models of SVD with double recentering.

- There are no clear relation between column recentering and row recentering. Neither do between no recentering and double recentering.

- It provides more insights to do all types of recentering at the exploration step.

Scree plot might help



What does the "best" mean?

- What kind of criterion should be used?

  – Best approximation?

  SVD with no recentering is always the best

  – Best interpretation?

  Provide more insights? How to find the best one?

  These problems are still under exploration

Summary

- SVD and PCA
- SVD surface plots
- SVD rotation movie
- SVD curve movie
- Matlab codes, movies and plots are online

Future work

- R package
- MATLAB package of SVD visualizations, combining our methods with other methods
- Other stuff related to SVD