
# Eigen Privy: Spectral Embedding of Documents and Topics

David J. Marchette

May 18, 2018

Acknowledgment: This work funded in part by the NSWC

Naval Innovative Science and Engineering (NISE) program.

NSWCDD-PN-18-00197 Distribution A: Approved for Public Release

# Topics

# Topics

# A Little History

- The Automated Serendipity (our name) project asked: Can one automate a DARPA program manager?
- That is: can one automate the matching of experts in disparate domains to produce innovations and breakthroughs?
- The team included Ed Wegman, Carey Priebe, Ronald Coiffman, Jeff Solka, Anna Tsao, DJM, and others.
- The main results were:
  - Methods to process documents, and produce a graph (as in manifold discovery).
  - Methods to produce an embedding of the documents (as in manifold discovery – spectral embedding).
- This led to a number of interesting ways to process text, and analyze networks.

# Spectral Embedding/Clustering

- Given a simple graph: $G = (V, E)$ (possibly directed).
- The goal is to analyze the network; in particular, can one cluster the vertices?
- Spectral Clustering is one way to cluster the vertices:
  - Take the adjacency matrix $A$ (or the Laplacian $D - A$, where $D$ is the diagonal matrix of degrees).
  - Possibly do something to this matrix:
    - Scale by $D^{-1/2}$.
    - Augment by adding a diagonal matrix, such as $D/(|V| - 1)$.
  - Embed via eigen (singular) vectors.
  - Cluster.

# Spectral Embedding Theory

- If the graph is a stochastic block model (SBM):
  - There are $g$ groups (clusters, communities) – I'll abuse notation and think of $g$ as a map from a vertex to a group as well as an indicator of the number of groups.
  - The probability of an edge between vertices depends only on the groups:

$$p_{ij} = P_{g(i)g(j)}$$

  - The edges are drawn independently.
- Then the embedding is a mixture of normals – asymptotically and some fiddly-bits, but that's the basic story.

# Spectral Clustering Questions

- Which should one use: Adjacency or Laplacian?
- Should one normalize by $D^{-1/2}$?
- What clustering method should one use?

# Spectral Clustering Questions

- Which should one use: Adjacency or Laplacian?
- Should one normalize by $D^{-1/2}$?
- What clustering method should one use?
- For the answers to these and many more questions, see Carey's talk.
- Instead, consider the following:
  - Each connected component of the graph needs to be addressed separately.
  - Or, one needs a way to extend the embedding "across the divide" between the components.
- First, let's take a detour into text analysis.

# Topics

# Quantitative Horizon Scanning

- Hypothesis: scientific breakthroughs happen when someone finds a way to combine information from two disparate fields – synthesizing information across fields to produce a novel idea.
- While this is not universally true, it is clearly anecdotally true.
- Can one find a way to detect potential syntheses prior to the breakthrough, particularly if someone else is about to make a breakthrough. We call this "mitigating technilogical surprise".
- Priebe et al. developed a log-odds ratio approach to detect when organization A is "poised" to merge two disciplines before organization B.
- This ratio depends on the concept of "priviness".

# b-Privy

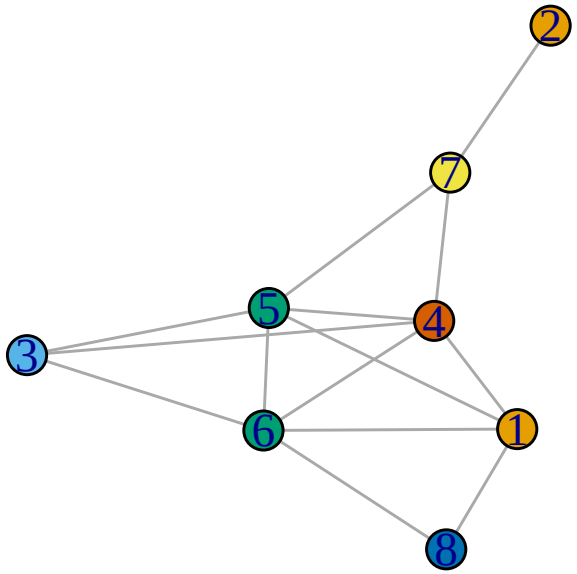
Set up:

- Each document has a set of topics associated to the document.
- Each author thus inherits the topics of the documents s/he (co)authored.
- From a collection of documents, one can construct the co-authorship graph:
  - Vertices are authors (people).
  - An edge between two authors indicates they were co-authors on at least one paper.

Priviness:

- An author is:
  - 0-privy to the topics they have written about.
  - 1-privy to those topics their co-authors have written about.
  - Etc.

# Computing b-Privy

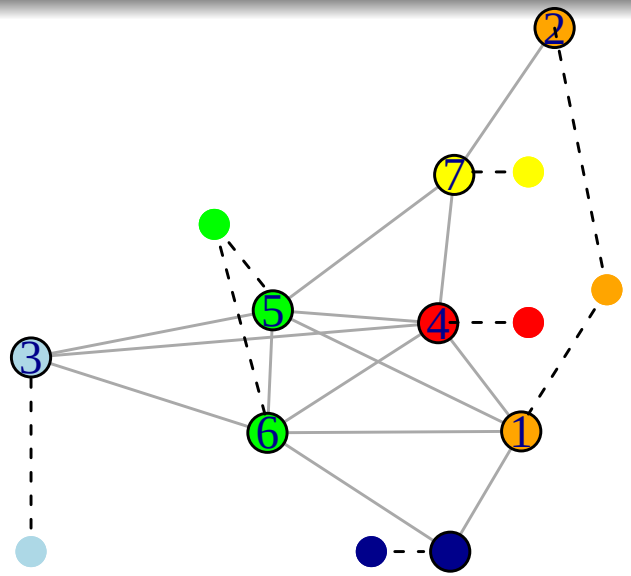



In the figure, each author has published on a single topic (node color). Author 1 is:

- 0-privy to orange.
- 1-privy to green, red and dark blue.
- 2-privy to light blue and yellow.

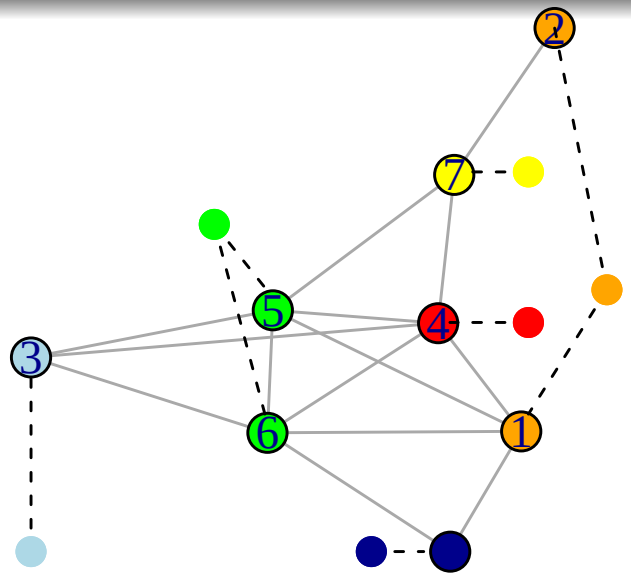
Author 2 is 4-privy to dark blue.

# Computing b-Privy




- Add new nodes, one per topic.
- Connect each author to the topics associated with the author.
- An author's "priviness" to a topic is the shortest path distance to the topic...almost...

# Computing b-Privy




- Add new nodes, one per topic.
- Connect each author to the topics associated with the author.
- An author's "priviness" to a topic is the shortest path distance to the topic...almost...
- To ensure the path does not "take a shortcut" through another topic, the original graph is treated as a bidirected graph, and the edges from the topics to the authors are directed.

# SBM Viewpoint

- Augmenting the graph can be thought of as adding new blocks of singleton vertices.
- This modifies the model a bit, since the edge probabilities from these blocks may not follow the block model.
- I hypothesize that the augmentation does one of the following to the mixture model:
  - Combine groups that otherwise were disconnected.
  - Split groups according to their connections to the topics.
  - Add a bit of noise (variance) if the clusters are independent of the topics.
  - Some interesting combination of the above.

# b-Privy Observations

- Because the topic–author edges are uni-directional, the b-privy calculations via the shortest path algorithm is correct.
- The co-authorship graph (or citation network) tends to be disconnected.
- The augmented graph tends to be much more connected – often with just a few tiny disconnected components (this statement depends very much on the topic granularity).

# b-Privy Observations

- One can now use the augmented graph in the spectral embedding.
- Note that this approach works for any categorical variable that is associated with (a subset of) the authors.
- Whether it makes sense to do this graph augmentation depends on the data and on the inference one wishes to make.
- We can control the effect of the augmentation by weighting the topic–author edges.
- In principal, one could give each topic its own weight – if one wished to investigate the influence of the different topics.

# Topics

# The Eigen Privy Idea

- Given a graph (directed or undirected) with categorical vertex attributes.
- If the graph is undirected, convert to directed graph by bi-directing the edges.
- Add nodes for each observed value of the categorical attribute(s).
- Add directed edges from these new nodes to each original vertex with that attribute value.
- Spectral embed the resulting graph.

# SBM Viewpoint

- Thinking of this as a stochastic block model:
- The original model is:

$$P[i \to j] = P[g(i) \to g(j)]$$

for $i, j \in V(G)$.

- The new model is:

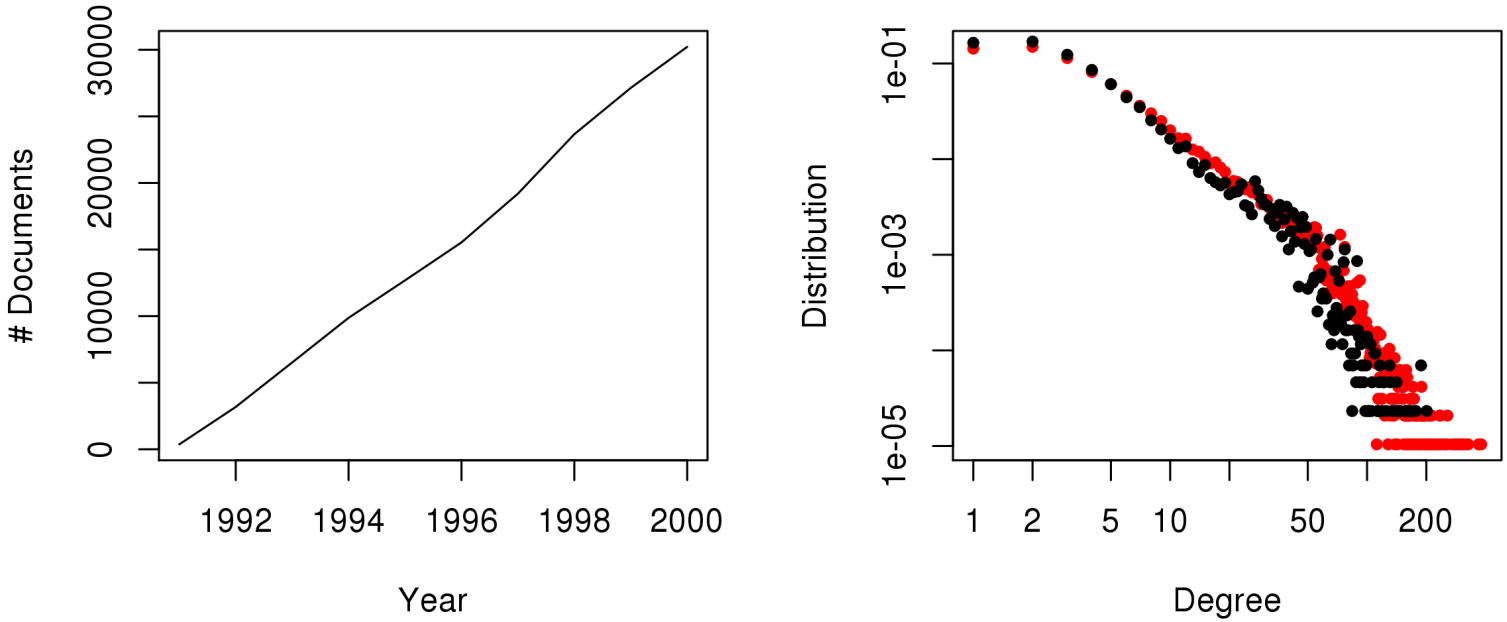
$$P[i \to j] = \begin{cases} P[g'(i) \to g'(j)] & i, j \in V(G) \\ p_{ij} & i \notin V, j \in V(G) \\ 0 & j \notin V(G) \end{cases}$$

- It is important to note that the groups may change with the augmentation.

# Example: arXiv

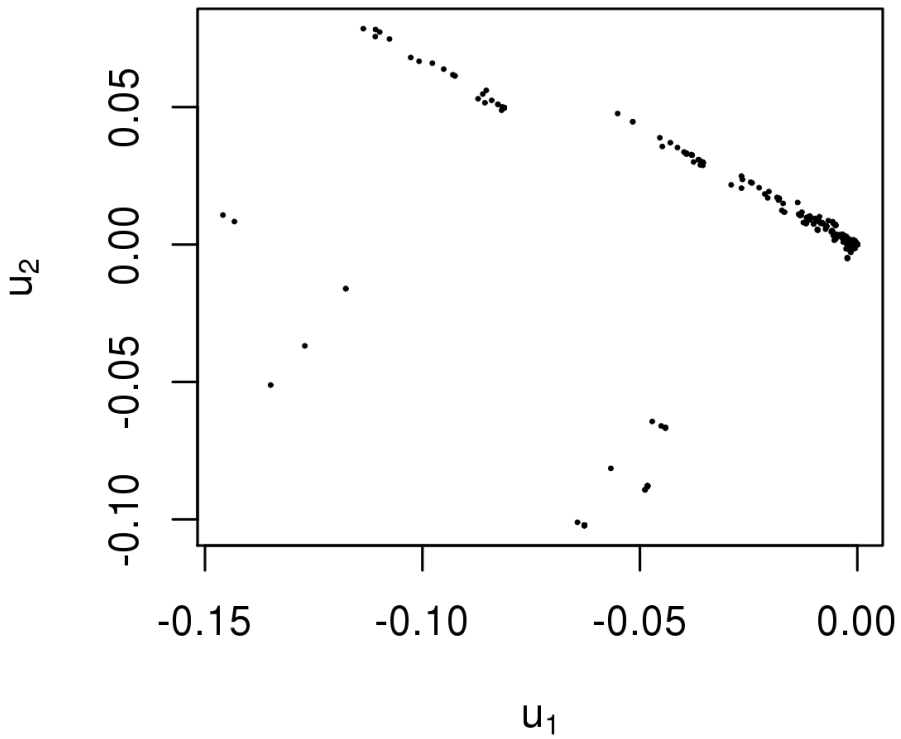
- A data set of arXiv documents from 1991 to 2000 was collected.
- The co-authorship graph was constructed, and the primary category was used as the topic for each document.
- No author dissambiguation was attempted – with all that implies.
- We'll focus on the articles published in the year 2000.

# arXiv Data

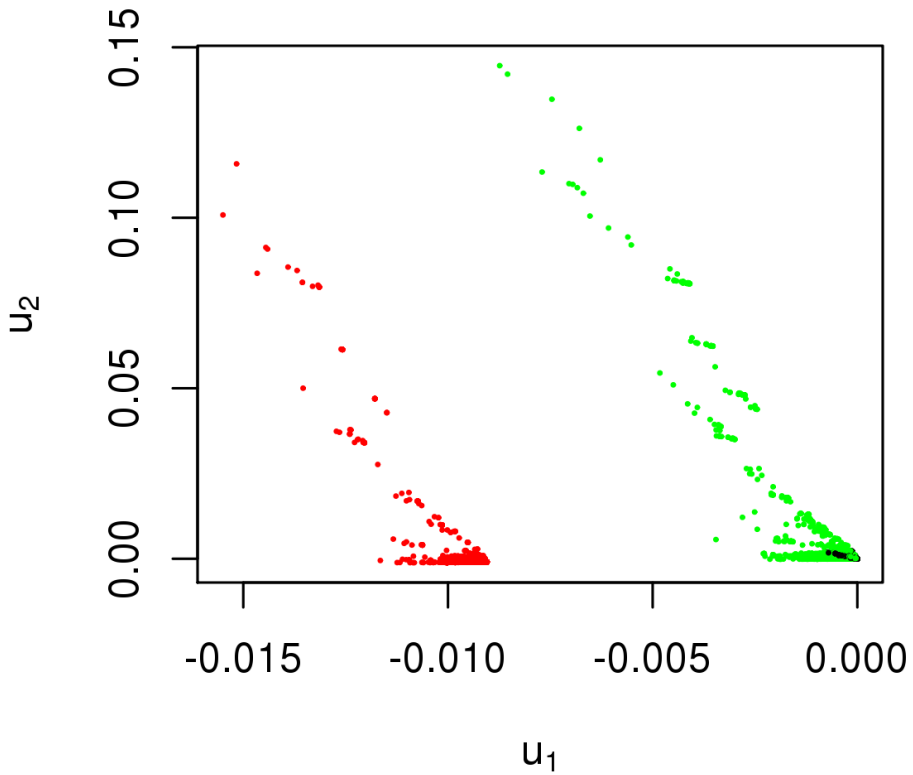



Number of documents in the arXiv dataset by year, degree distributions (black=2000, red=1999–2000).

# arXiv Data 2000 – Spectral Embedding



# arXiv Data 2000 – Spectral Embedding

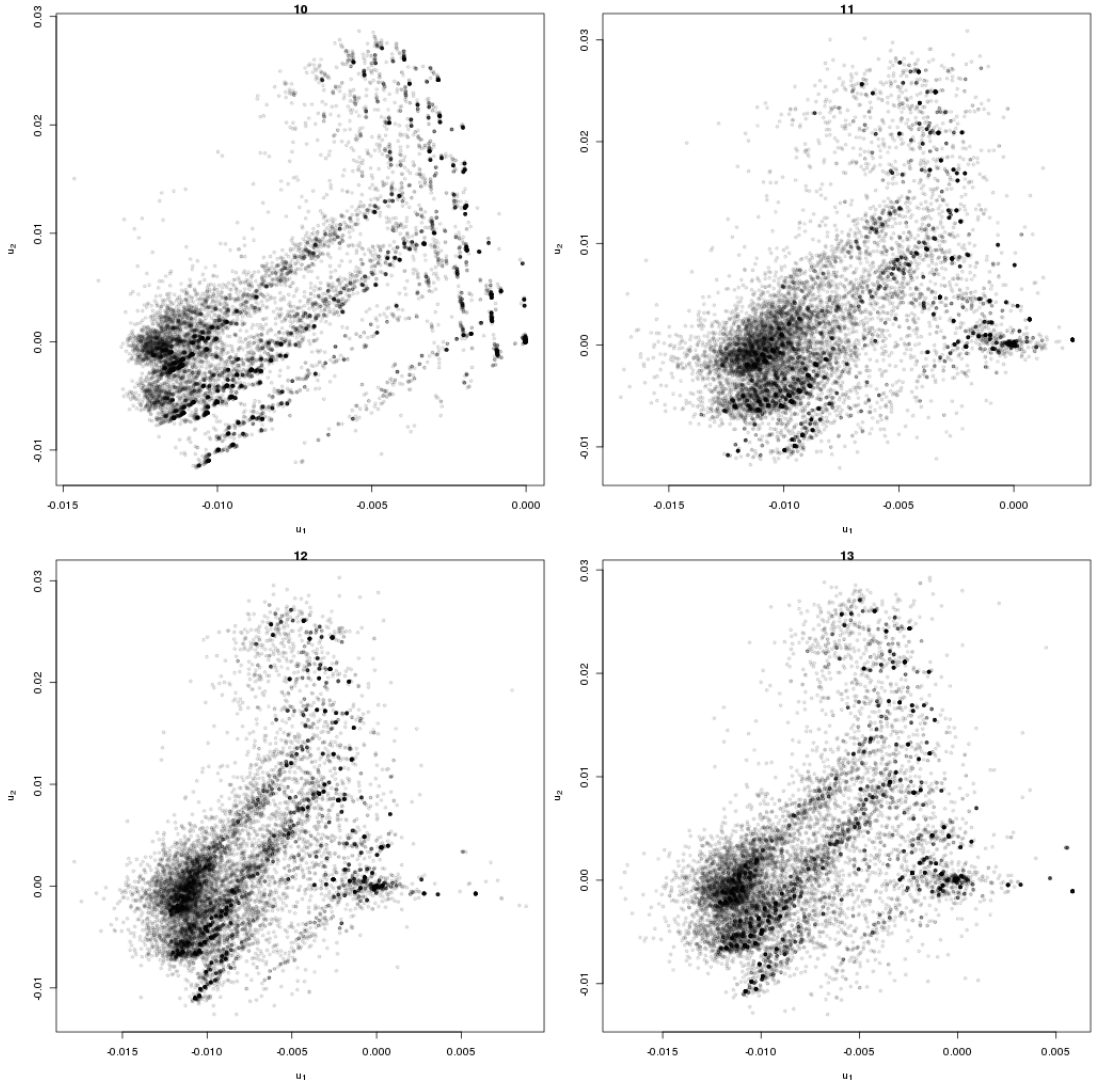



Red: Astronomers, Green: 1-privy to astronomy, Black: others.

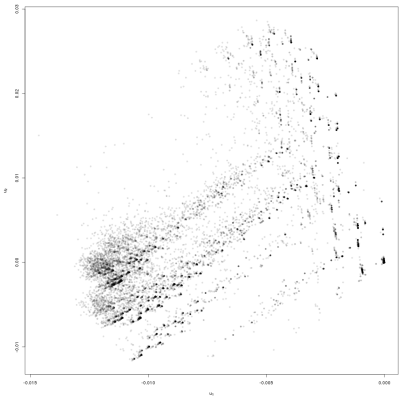
# Example: LANL Network Flows

- A data set of network flows – connections between machines – available from `https://csr.lanl.gov/data/2017.html`.
- Consider one day of TCP connections.
- Nodes are computers, attributes are common services – destination ports.
- Augmenting reduces from 10 connected components to 2 – not a significant change – only about 30 computers not originally in the largest component.

# Graph Embedding – 4 Hourly Graphs

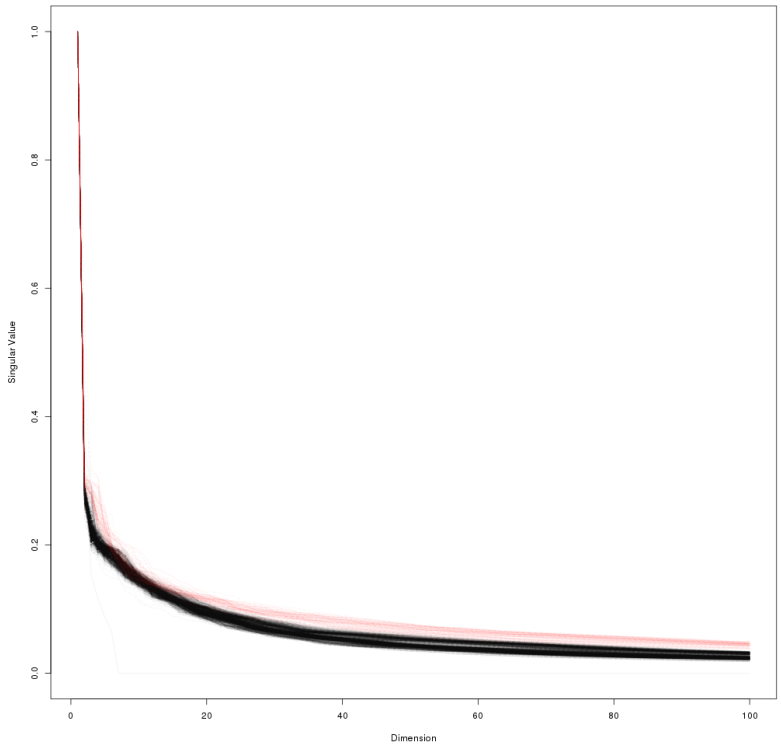


# Comments



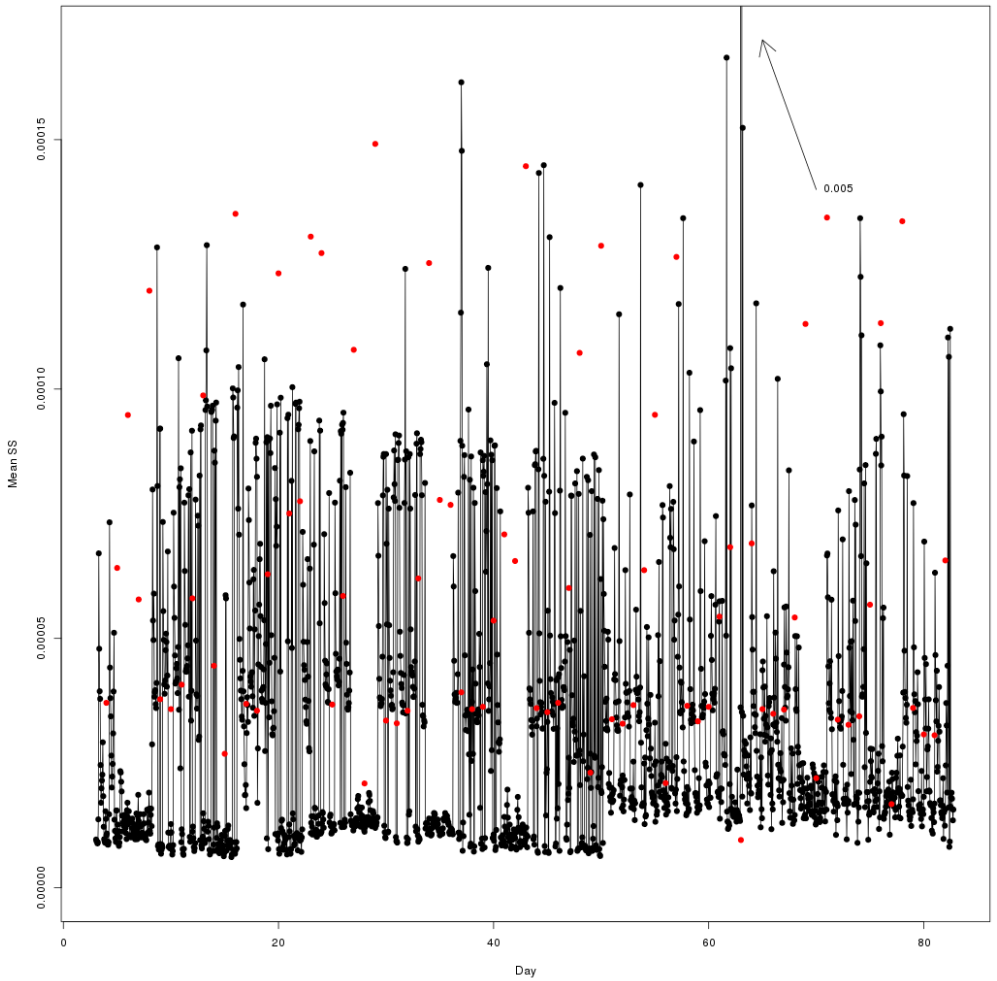
- This is only a 2d embedding, so take with a grain of salt.
- Clear structure – not quite a Gaussian mixture?
- Structure is stable from hour to hour (for this time period).
- We can use this to construct measures of overal "normal behavior".

# What Should $d$ Be?



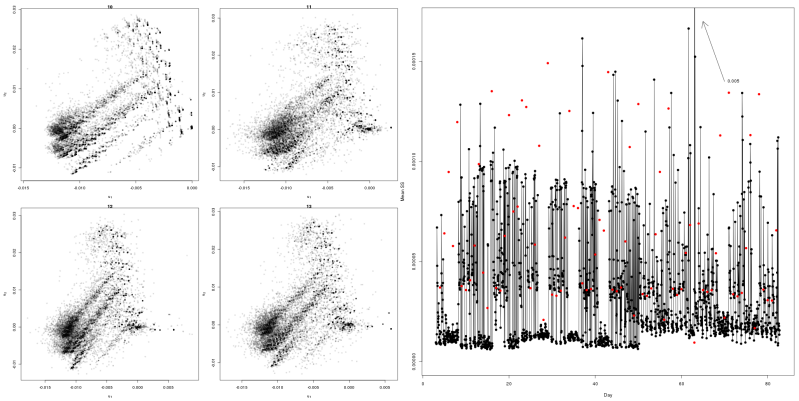

- Scree plot for 80 days. Hourly (black) daily (red).
- Elbow seems to be at the very low end.
- Maybe $d$ should be around 3 for hourly data, 3–7 for daily?

# Sum of Squared Differences




Emedding to $\mathbb{R}^3$, Red: day, Black: Hour.

# Thoughts and Future Work



- The sum-of-squares residual from the Procrustes allows for computer-specific analysis.
- The Procrustes is only computed on the machines in common. What about a global change measure?
  - Density-based approach:
    - Kernel densities computed on the two projections (if low dimensional).
    - Compute distances between the densities: KL-divergence, Hellinger, Integrated Squared (Marron and Wand 92).
  - Is there an appropriate mixture model for these data?