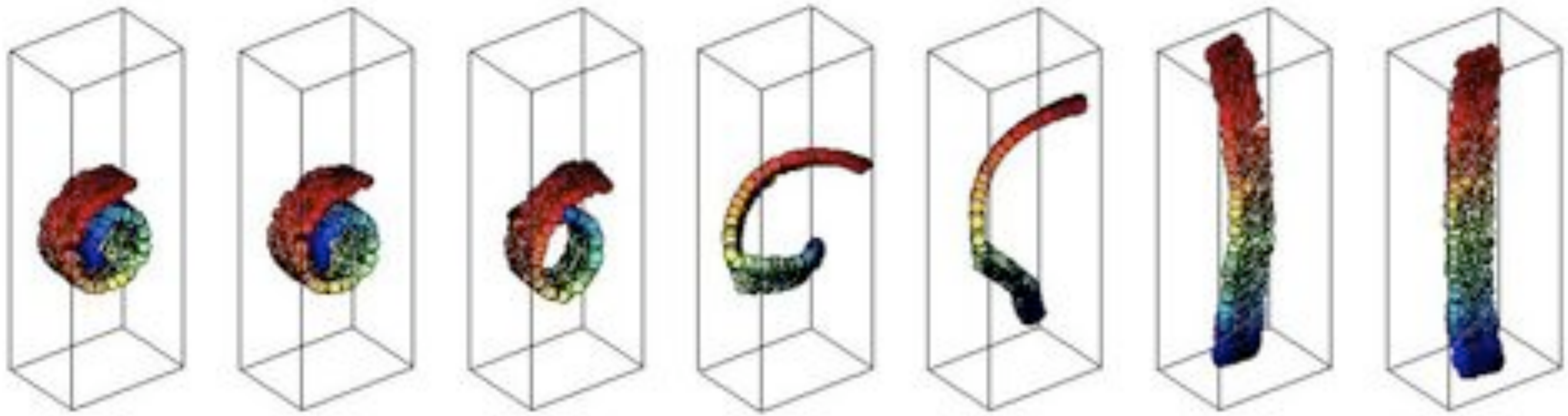
Semidefinite Embedding

Visualizing Folksonomy

Blake Shaw
bs2018@columbia.edu

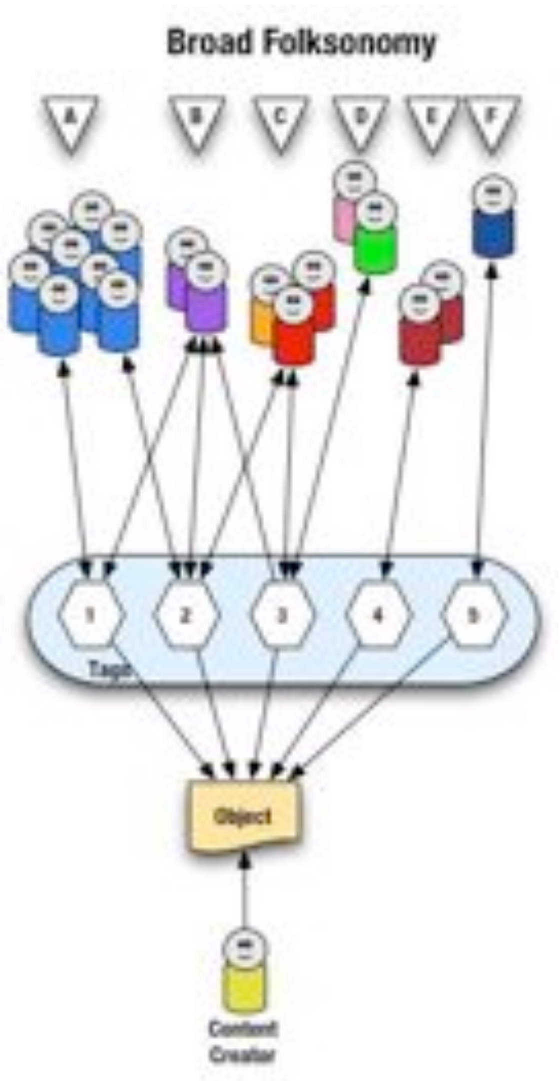
What is SDE?

- An algorithm to find a low dimensional nonlinear manifold that best fits a high dimensional data set.
- Formulates the problem to be solved by a semidefinite programming package
- “Unfolds” the data while trying maximize pairwise distances



Visualizing Folksonomy

- The del.icio.us service is a social bookmarking tool where users tag links with descriptive keywords.
- The goal is to visualize the relationships between these tags

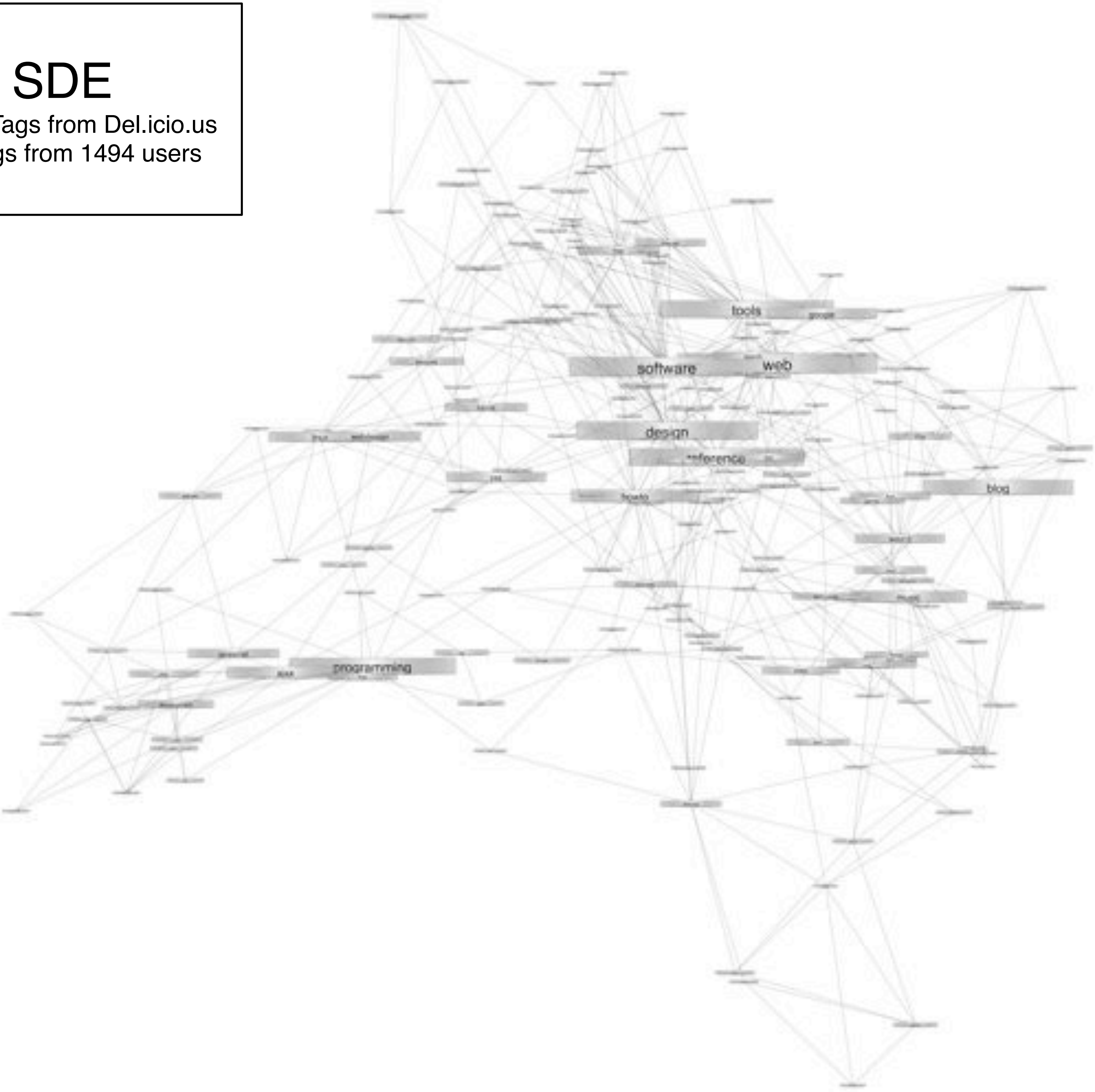


Project Goals

- Build a simple SDE package
- Apply this technique to visualizing tags
- Investigate heuristics for picking the best parameters

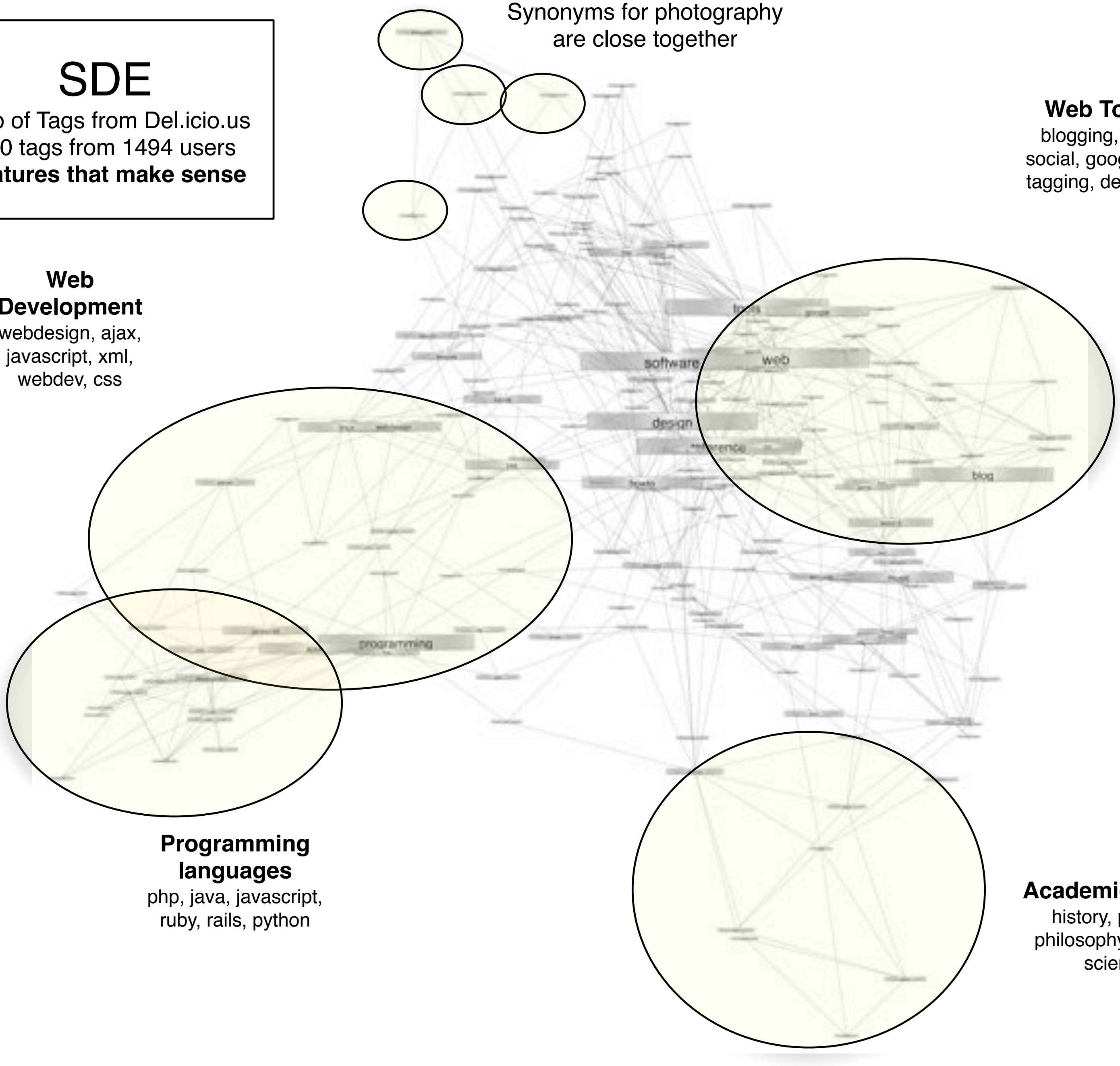
SDE

Map of Tags from Del.icio.us
200 tags from 1494 users



SDE

Map of Tags from Del.icio.us
200 tags from 1494 users
Features that make sense



Synonyms for photography are close together

Web Topics
blogging, blogs, social, google tags, tagging, deli.icio.us

Web Development
webdesign, ajax, javascript, xml, webdev, css

Programming languages
php, java, javascript, ruby, rails, python

Academic Topics
history, politics, philosophy, culture, science

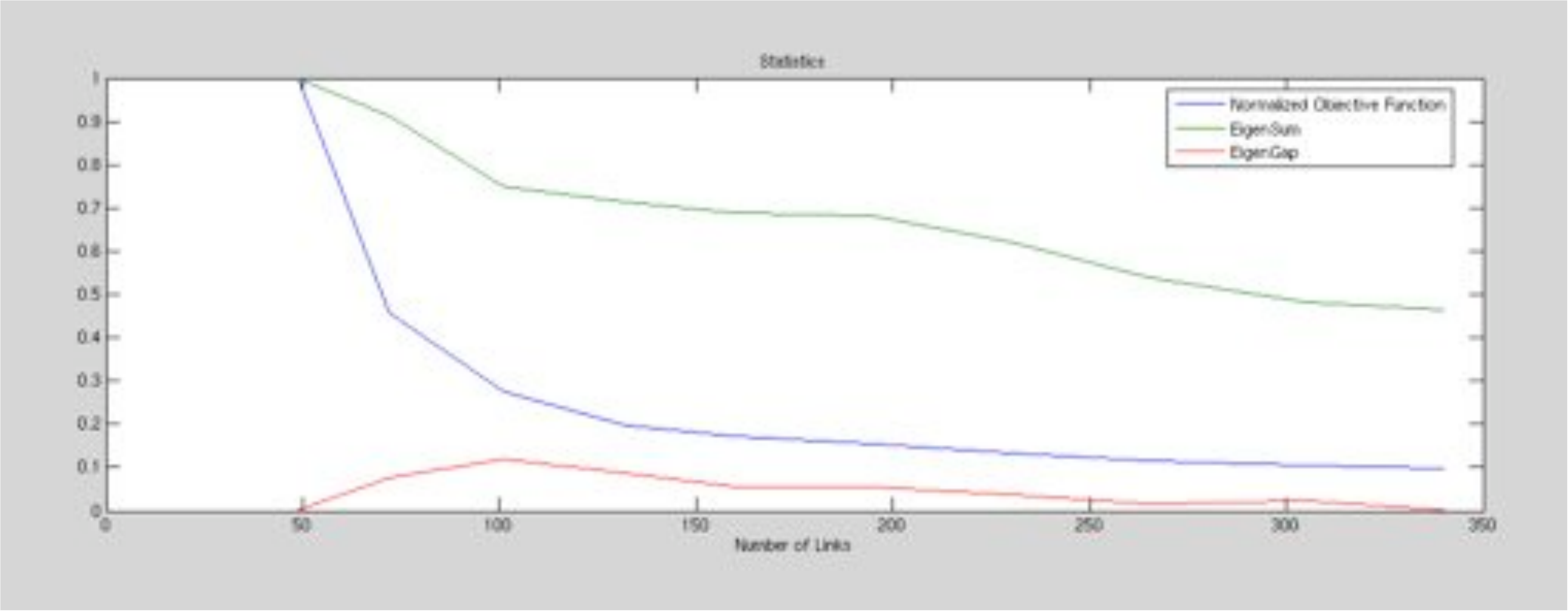
Matlab Demo

SDE Parameters

- Distance Metric
  - Can be Euclidean, KL Divergence, or Kernels
- Specify connectivity matrix
  - The algorithm assumes that only local distances can be trusted
  - Typically uses k-nearest neighbors

Choosing K

Objective Function, EigenSum, EigenGap



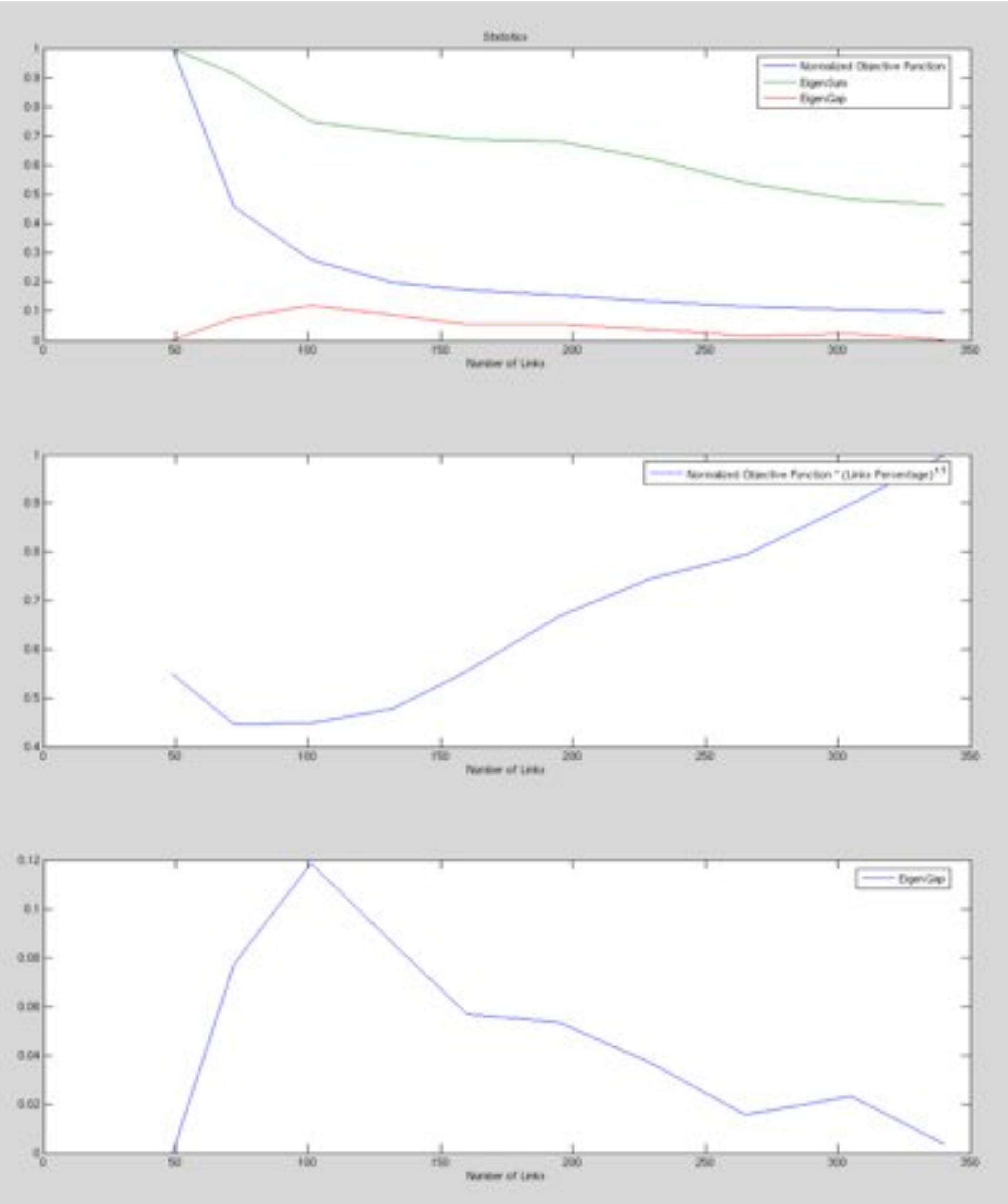
N = 50, K = {0, 1 ... 9}

Initially uses minimum spanning tree

Choosing K

N = 50, K = {0, 1 ... 9}

Initially uses minimum spanning tree



Need to find a balance between adding complexity and reducing the quality of the embedding

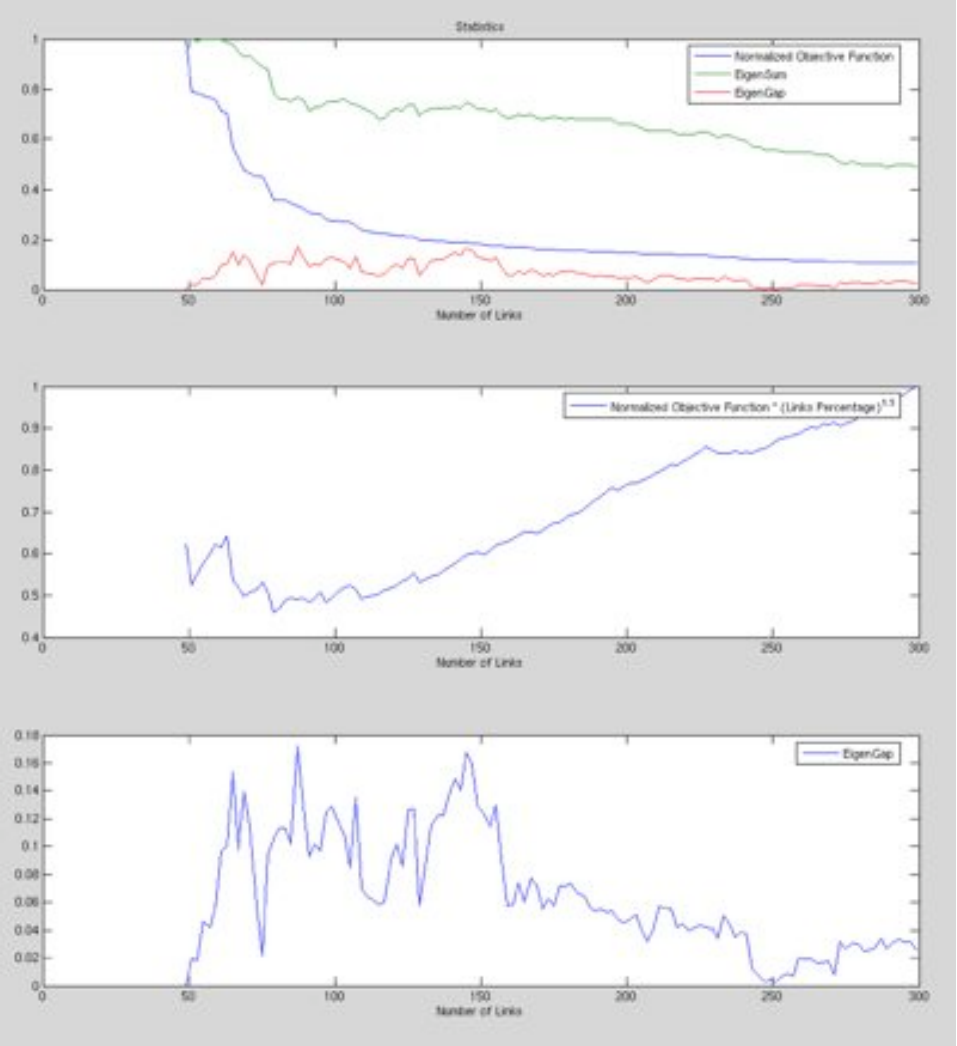
A heuristic that pinpoints the drastic change in the objective function

EigenGap -- a good measure of how well the data fits in a lower dimension

Variations

- Adding links incrementally
  - Overall best links first
  - Local best links first
- Higher degree nodes get more links
- Other datasets
  - A well known cancer dataset

Choosing K

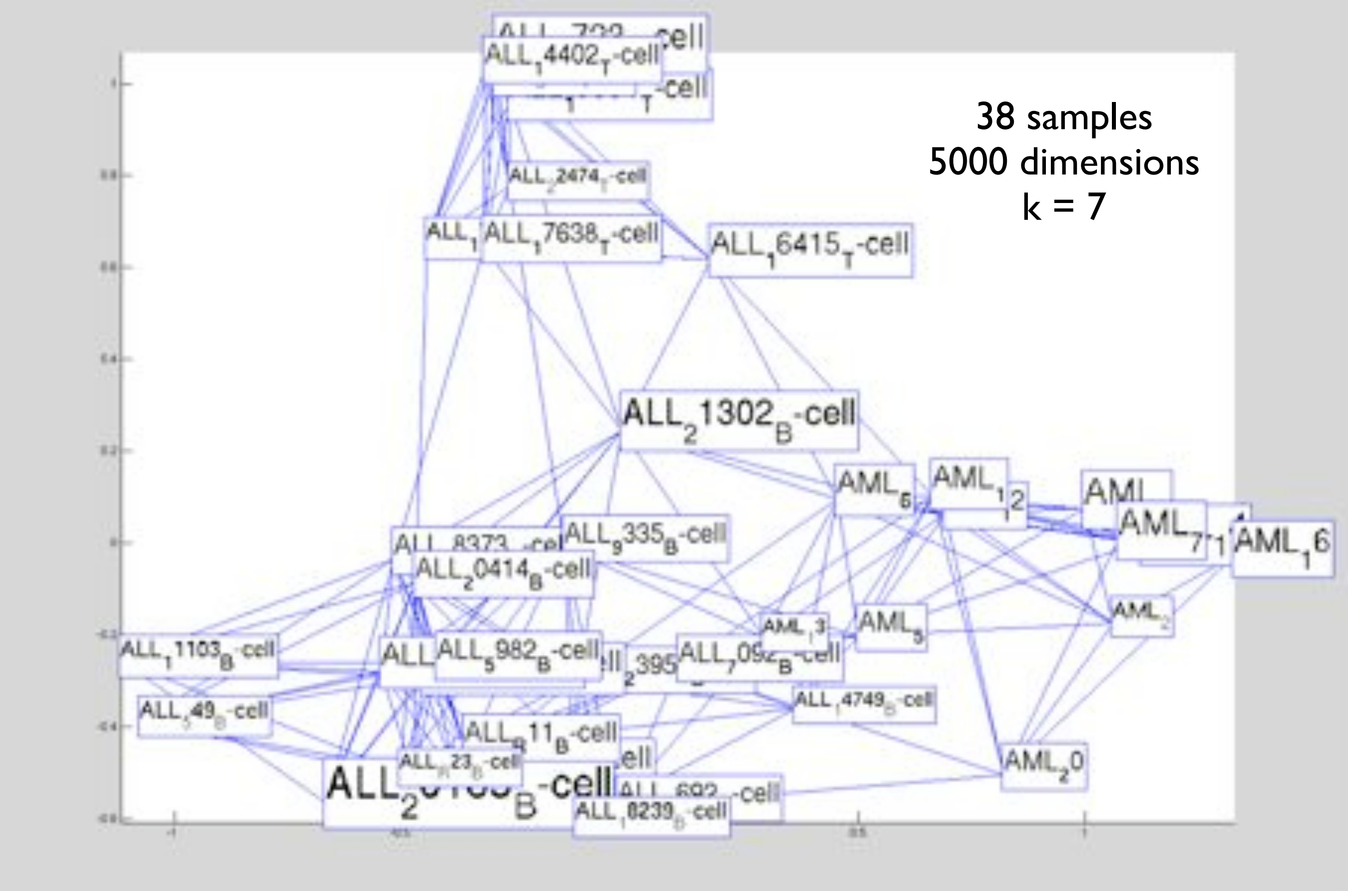


N = 50, Number of links = {0 - 300}

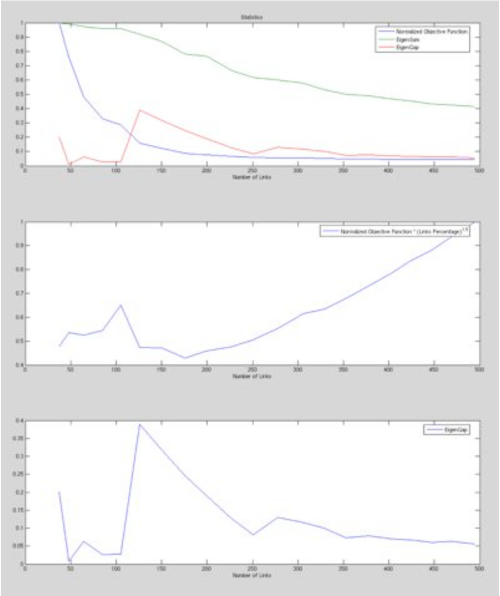
Like K-nearest but incrementally adds links

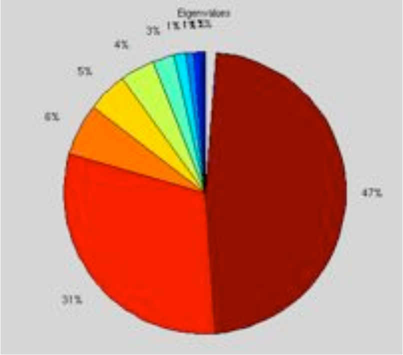
Leukemia Dataset

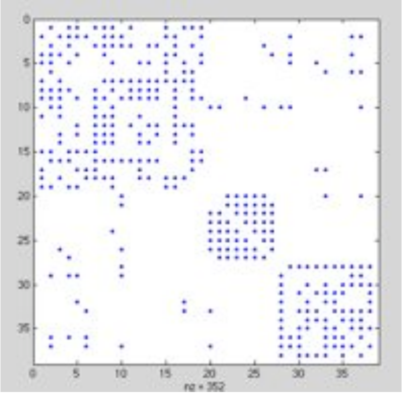
Golub -- microarray data



38 samples
5000 dimensions
k = 7







Conclusions

- Variations of the connectivity matrix can drastically change the low dimensional embedding
- We need better metrics to assess the quality of a connectivity matrix.

Future Work

- Pick best connectivity matrix through a more graph-oriented algorithm. Before embedding with SDE.
    - For instance, prune edges while trying to maintain certain properties: clustering coefficient, degree centrality, average path length, etc...
- Provide a more rigorous mathematical basis for comparing embeddings created by different distance metrics, connectivity regimes, etc...