Status Update

Automatic Parameter Tuning

- Threshold setting for MI graph
  - Simple heuristic: Set threshold to be 1 standard deviation away from the minimum weight till average node degree ~2 (graph is more scale free).
- Amount to discount between joins
  - Not a lot of work has been done
- Size of resulting clusters
  - Set arbitrary threshold on maximum number of clusters (currently 50) and set the minimum number of elements of per cluster accordingly.

MI discounting

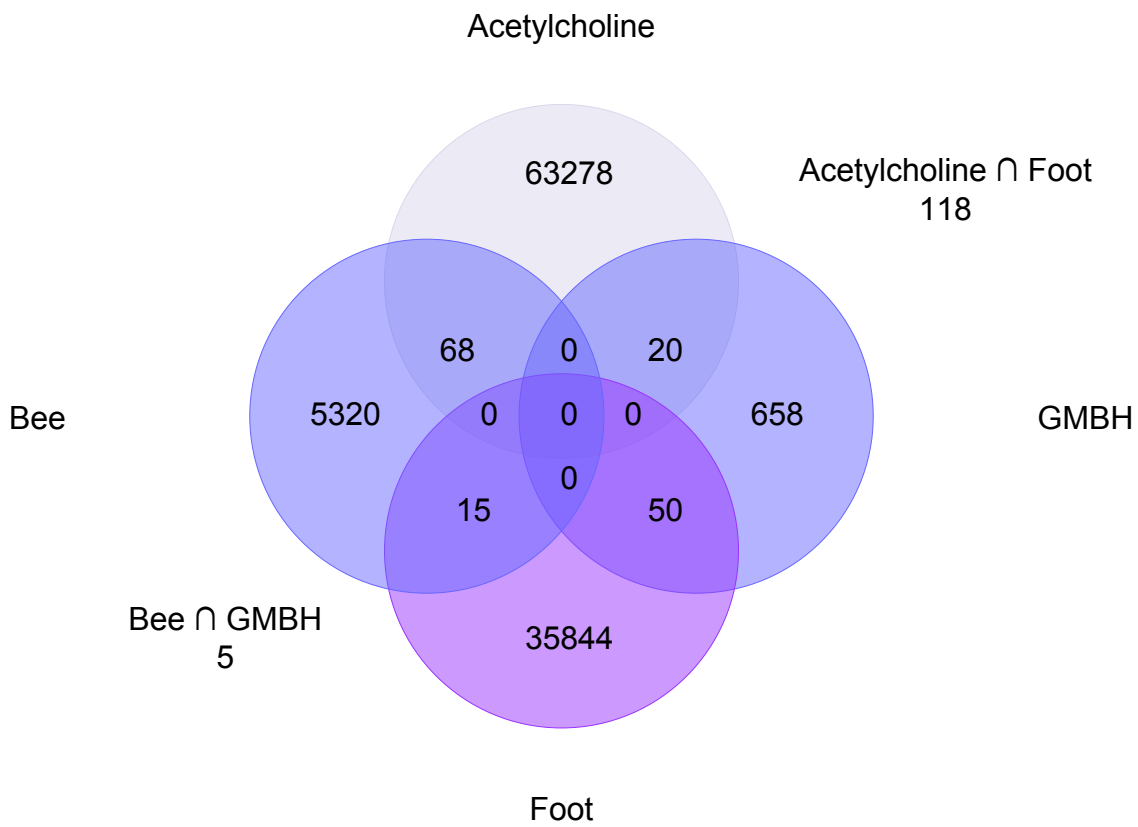
Evaluation

- Datasets
  - Artificial dataset
    - Disparate subsets of medline
  - Tagged corpus
    - Reuters 1997 corpus w/ topic categories
- Measures
  - Fmeasure
  - Entropy
  - Intra/Inter cluster similarity/distance

Example: Artificial Disparate Collection

- Subset of Medline:
  - Acetylcholine 64K documents
  - GMBH 700 documents
  - Bee 5K documents
  - Foot 36K documents

Collection overlap



Artificial Collection

- Results:
  - http://beespace.cs.uiuc.edu/~chee/artificial_results.txt