



The image features seven circles arranged in two rows. The top row contains three circles: an empty circle with a light purple outline on the left, and two solid light purple circles on the right. The bottom row contains four circles: two solid light purple circles on the left and two empty circles with light purple outlines on the right. The text 'Status Update' is centered horizontally between the two rows, overlapping the two solid circles in the top row and the two empty circles in the bottom row.

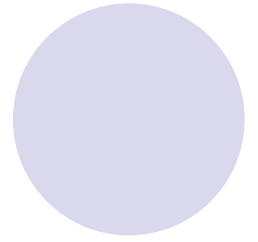
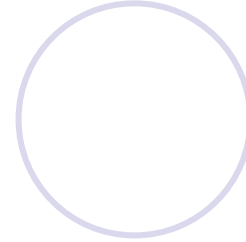
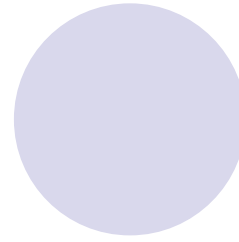
Status Update

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

A decorative graphic at the top of the slide consists of two rows of circles. The first row has a solid light purple circle on the left and an outlined light purple circle on the right. The second row has a solid light purple circle on the left, an outlined light purple circle in the middle, and a solid light purple circle on the right.

● Datasets

○ Artificial dataset

- Disparate subsets of medline

○ Tagged corpus

- Reuters 1997 corpus w/ topic categories

● Measures

○ Fmeasure

○ Entropy

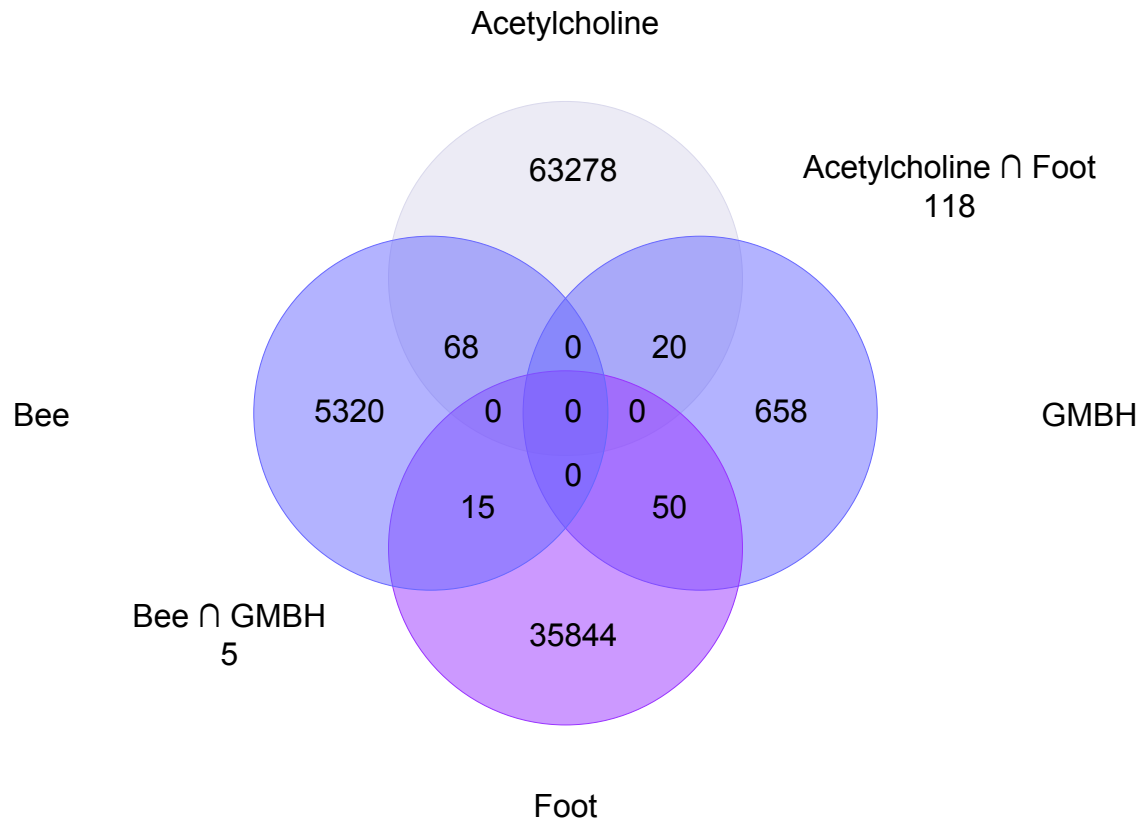
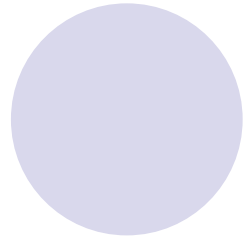
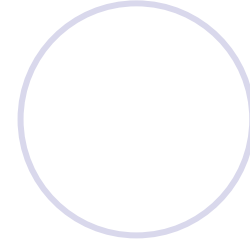
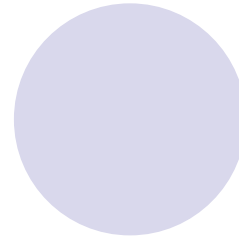
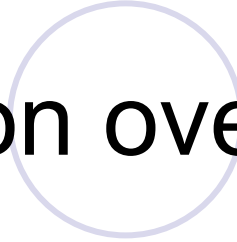
○ Intra/Inter cluster similarity/distance

Example: Artificial Disparate Collection

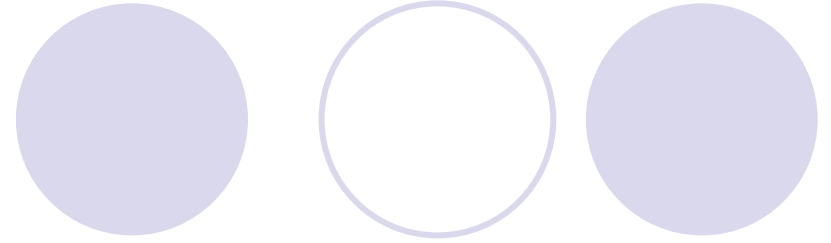
- Subset of Medline:

- 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