

Intersection Graphs for Text Analysis

Elizabeth Leeds

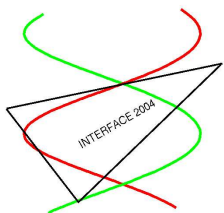
leedsem@nswc.navy.mil

David Marchette

marchettedj@nswc.navy.mil

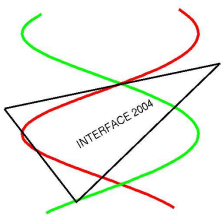
Naval Surface Warfare Center

Code B10



Overview

- bag-of-words approach to document encoding
 - word weighting by mutual information
 - only “important” words are kept
- intersection graphs are used to analyze document relationships
 - each document is a vertex
 - an edge exists between two documents if they share important words



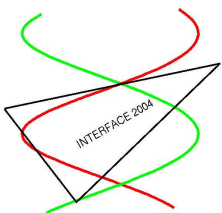
Mutual Information

Let $c_{w,d}$ be the number of times that the word w has occurred in the document d and let $N_{\mathcal{S}}$ be the total number of words (counting duplicates) in the corpus \mathcal{S} . Let $f_{w,d} = c_{w,d}/N_{\mathcal{S}}$. Then the mutual information between document d and word w is given by

$$m_{w,d}^{\mathcal{S}} = \log \left(\frac{f_{w,d}}{\sum_{z \in \mathcal{S}} f_{w,z} \sum_i f_{i,d}} \right) \quad (1)$$

Let N_d be the number of words (counting duplicates) in document d . Let $c_{w,\mathcal{S}}$ be the number of times that the word w appears in the corpus \mathcal{S} .

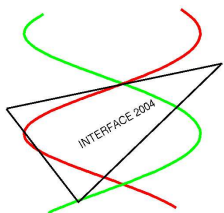
$$m_{w,d}^{\mathcal{S}} = \log \left(\frac{\frac{c_{w,d}}{N_d}}{\frac{c_{w,\mathcal{S}}}{N_{\mathcal{S}}}} \right)$$



Mutual Information - Summary

- $c_{w,d}$ - the number of times that the word w appears in the document d .
- $c_{w,\mathcal{S}}$ - the number of times that the word w appears in the corpus \mathcal{S} .
- N_d - the number of words (counting duplicates) in document d .
- $N_{\mathcal{S}}$ - the total number of words (counting duplicates) in the corpus \mathcal{S} .

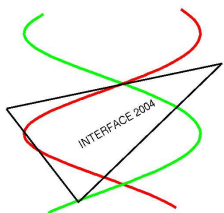
$$m_{w,d}^{\mathcal{S}} = \log \left(\frac{\frac{c_{w,d}}{N_d}}{\frac{c_{w,\mathcal{S}}}{N_{\mathcal{S}}}} \right) \quad (2)$$



Intersection Graphs and the KSS Random Intersection Graph

- G is an *intersection graph* if a set S_v can be assigned to each vertex $v \in V(G)$ so that $vw \in E(G)$ exactly when $S_v \cap S_w \neq \emptyset$.
- To define a *random intersection graph*, let $p \in [0, 1]$ and let $M = \{1, 2, \dots, m\}$. Define n random subsets $S_k, k = 1, \dots, n$ of the set M where each element of M is selected for the subset S_k with probability p . Then $G(n, m, p)$ is the intersection graph of the sets S_k .

Karonski, Scheinerman, Singer-Cohen, (1999) On Random Intersection Graphs: The Subgraph Problem. In *Combinatorics, Probability and Computing*, Vol 8, pp. 131-159.



Thresholding

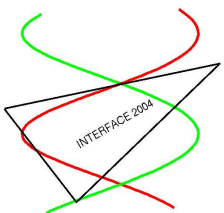
For each document (vertex) we have a set of words with each word assigned a weight.

- Let S_j be the set of words contained in document j
- Let $d_j = \{m_1, m_2, \dots, m_{|S_j|}\}$ be the ordered set containing the weights for each word in S_j .

Consider two types of thresholding:

$$t(m, \tau) = \begin{cases} 0 & \text{if } m < \tau \\ m & \text{if } m \geq \tau \end{cases}$$

$$T(m, \tau) = \begin{cases} 0 & \text{if } m < \tau \\ 1 & \text{if } m \geq \tau \end{cases}$$



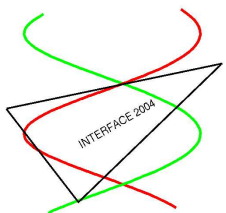
Defining Edges

- Under the KSS model, $v_i v_j \in E(G)$ if $S_i \cap S_j \neq \emptyset$.
- Modify this by taking $v_i v_j \in E(G)$ if:

$$|S_i \cap S_j| \geq k \text{ for some } k \in \mathbb{Z}^+$$

$$\frac{|S_i \cap S_j|}{\sqrt{|S_i||S_j|}} \geq q \in \mathbb{R}^+$$

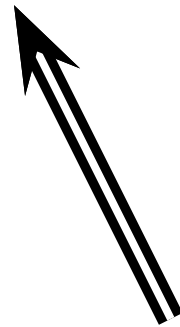
$$\frac{d_i \cdot d_j}{\|d_i\| \|d_j\|} \geq q \in \mathbb{R}^+$$



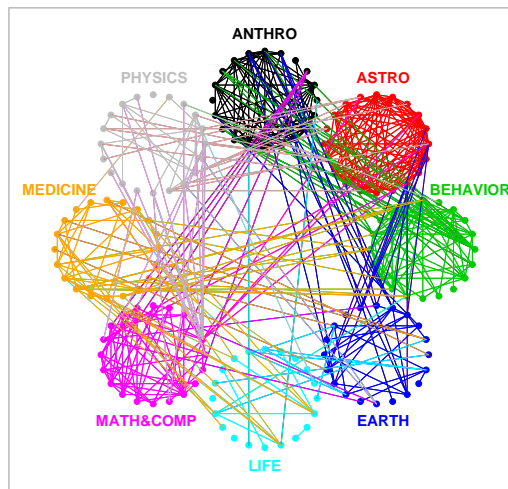
Procedure



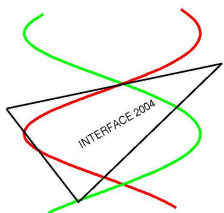
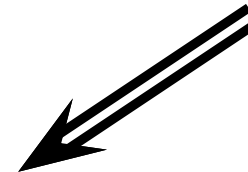
a	0.03
about	-0.26
abstract	4.22
accent	5.83
...	...
word	1.52
would	-0.26
year	0.50
young	2.79
yowlumni	5.83



Graph Size = 500
Mutual Information Threshold = 1

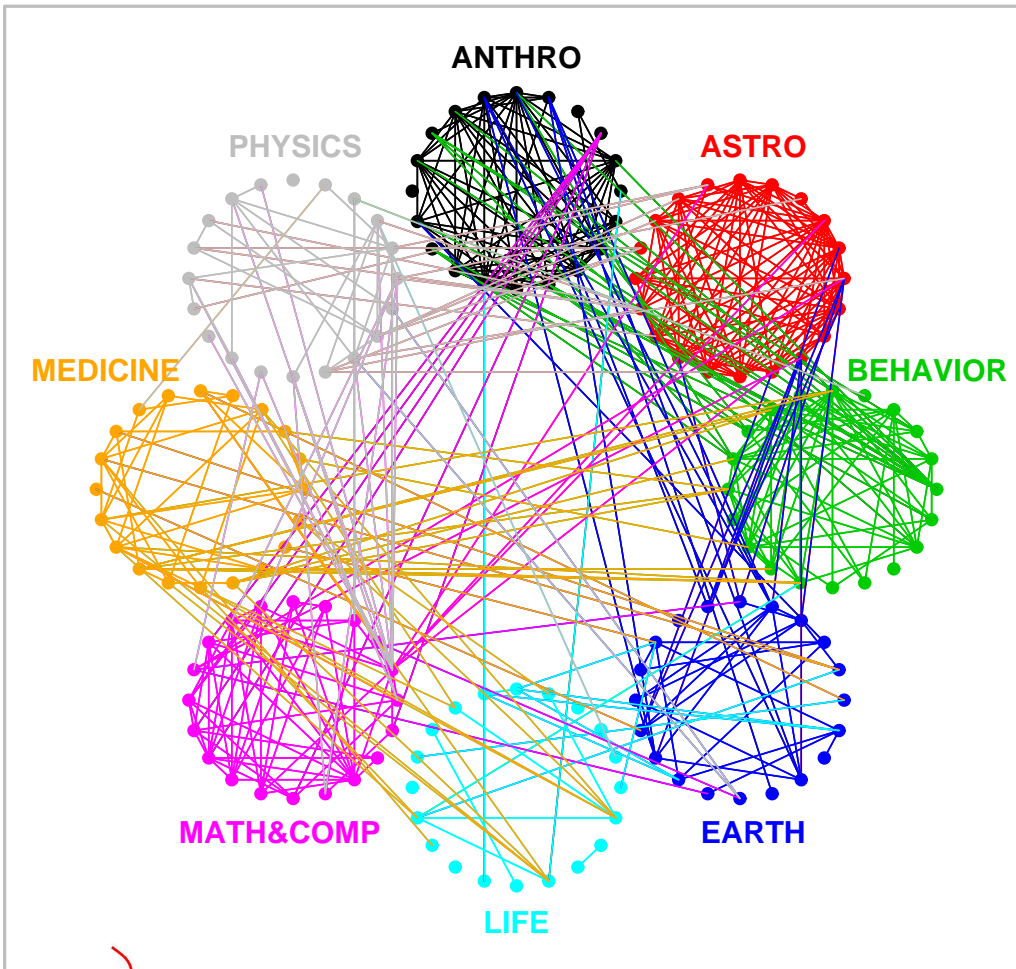


141 edges between classes



Intersection Graph

Graph Size = 500
Mutual Information Threshold = 1

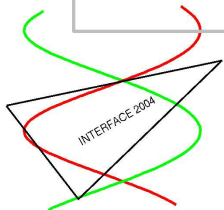


vertices are documents

threshold determines which words are important

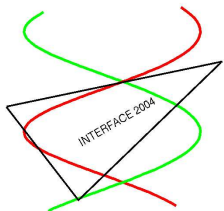
edge between documents that share important words

141 edges between classes

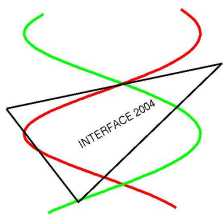
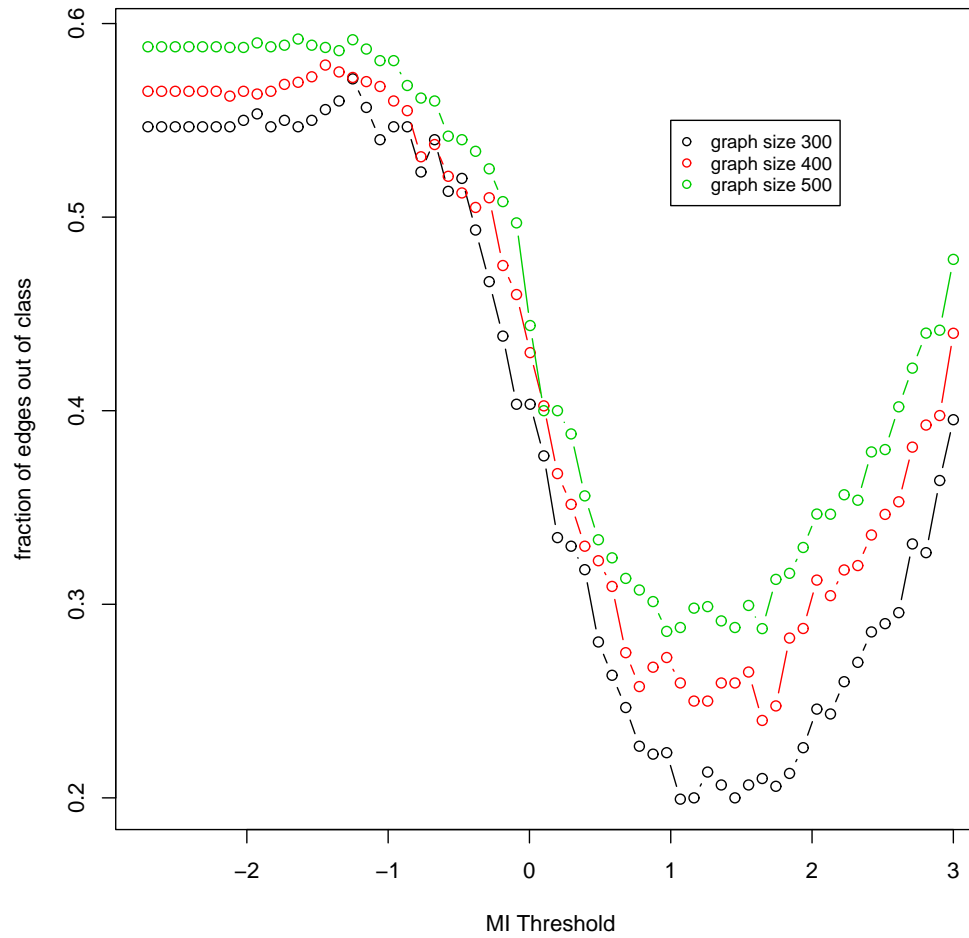


Mutual Information

- The weight is based on the frequency of the word in the document compared to the frequency of the word in other documents in the corpus
- Words that are important have large weights
- Throw out words with small weights
 - Reduces dimensionality
 - Reduces the noise
- What does "important" mean in terms of the mutual information?
 - Use graphs to select threshold value defining importance.
- This is different than the usual stopper list
 - Document/corpus dependent stopper list
 - Requires no knowledge of the language



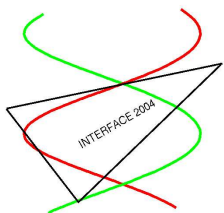
Using Mutual Information to Threshold



Adding Documents to the Corpus

- Add a new set of documents to the corpus.
- The weights on (importance of) the words in the original documents will change.
- What does the intersection graph tell us about this change?
- How can we use documents or sets of documents to force connections in the intersection graph?

Mathematically, a new set of documents changes the weight on a word by the same amount across all original documents.



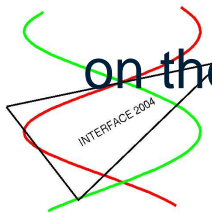
Adding Documents to the Corpus

Let the document d be in the corpus \mathcal{S} . Suppose we add a new set of documents, \mathcal{S}_1 , to \mathcal{S} and measure the change of $m_{w,d}$ under this change in corpus.

The change in the mutual information of word w in document d under the addition of the set of documents \mathcal{S}_1 is

$$\begin{aligned}\Delta_{m_{w,d}}^{\mathcal{S}_1} &= m_{w,d}^{\mathcal{S} \cup \mathcal{S}_1} - m_{w,d}^{\mathcal{S}} \\ &= \log \left(\frac{c_{w,d}}{N_d} \frac{N_{\mathcal{S} \cup \mathcal{S}_1}}{c_{w,\mathcal{S} \cup \mathcal{S}_1}} \right) - \log \left(\frac{c_{w,d}}{N_d} \frac{N_{\mathcal{S}}}{c_{w,\mathcal{S}}} \right) \\ &= \log \left(\frac{c_{w,\mathcal{S}}}{N_{\mathcal{S}}} \frac{N_{\mathcal{S} \cup \mathcal{S}_1}}{c_{w,\mathcal{S} \cup \mathcal{S}_1}} \right)\end{aligned}\tag{3}$$

The change in the mutual information for the word w does not depend on the document d .

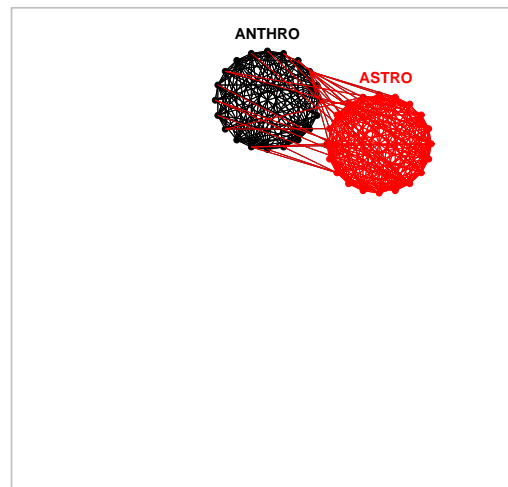


Adding Documents to the Corpus

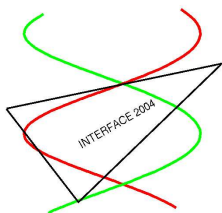
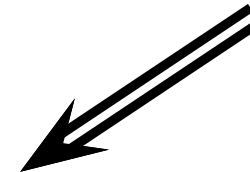


a	0.03
about	-0.26
abstract	4.22
accent	5.83
...	...
word	1.52
would	-0.26
year	0.50
young	2.79
yowlumni	5.83

Graph Size = 300
Mutual Information Threshold = 0.5



31 edges between classes

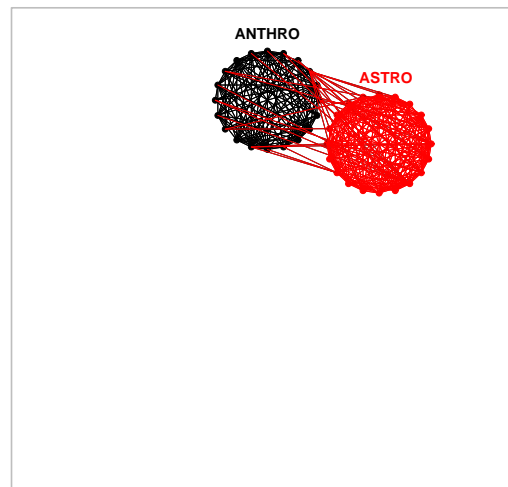


Adding Documents to the Corpus

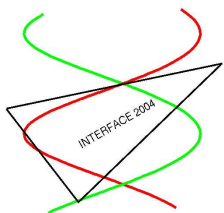
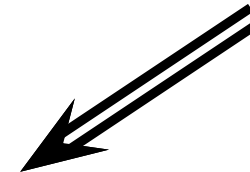


a	0.03
about	-0.26
abstract	4.22
accent	5.83
...	...
word	1.52
would	-0.26
year	0.50
young	2.79
yowlumni	5.83

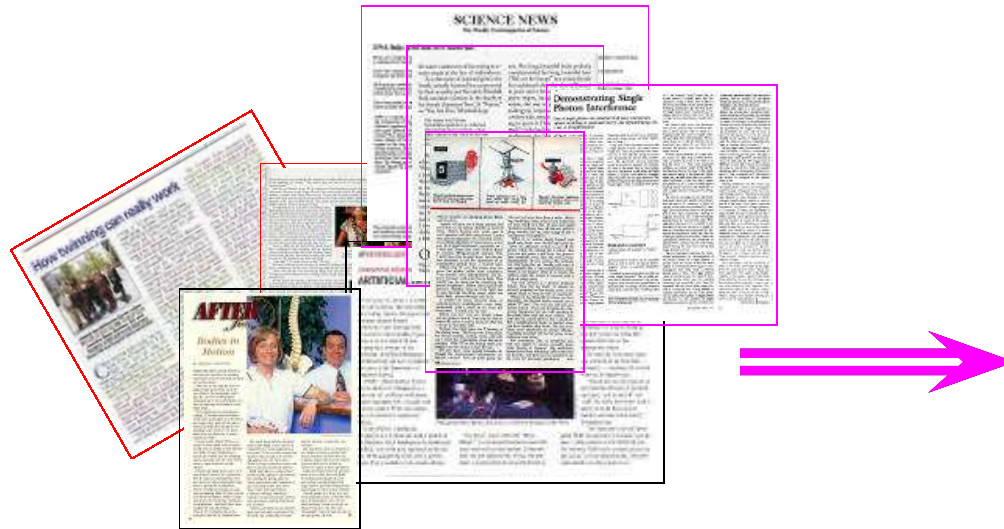
Graph Size = 300
Mutual Information Threshold = 0.5



31 edges between classes

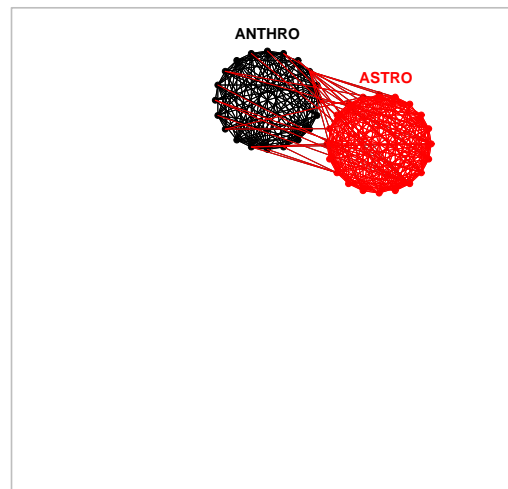


Adding Documents to the Corpus

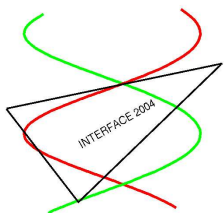


a	0.03	0.02
about	-0.26	-0.14
abstract	4.22	4.76
accent	5.83	6.15
...
word	1.52	4.23
would	-0.26	0.03
year	0.50	2.67
young	2.79	4.12
yowlumni	5.83	6.24

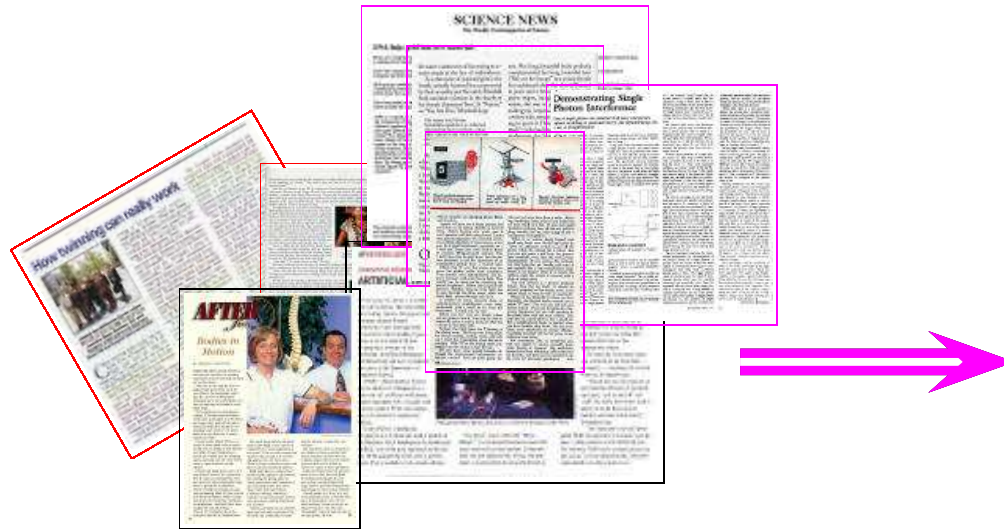
Graph Size = 300
Mutual Information Threshold = 0.5



31 edges between classes

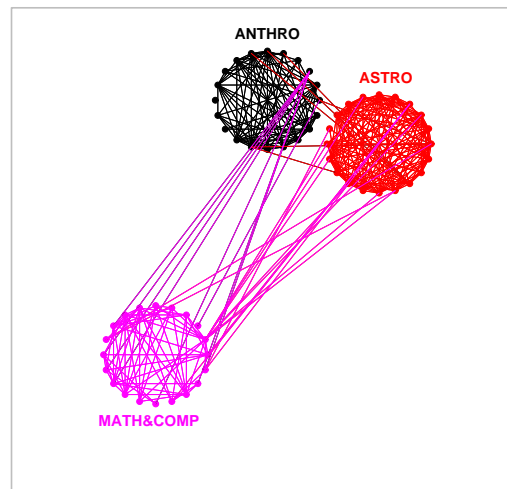


Adding Documents to the Corpus

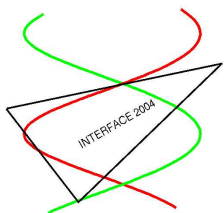


a	0.03	0.02
about	-0.26	-0.14
abstract	4.22	4.76
accent	5.83	6.15
...
word	1.52	4.23
would	-0.26	0.03
year	0.50	2.67
young	2.79	4.12
yowlumni	5.83	6.24

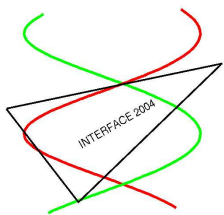
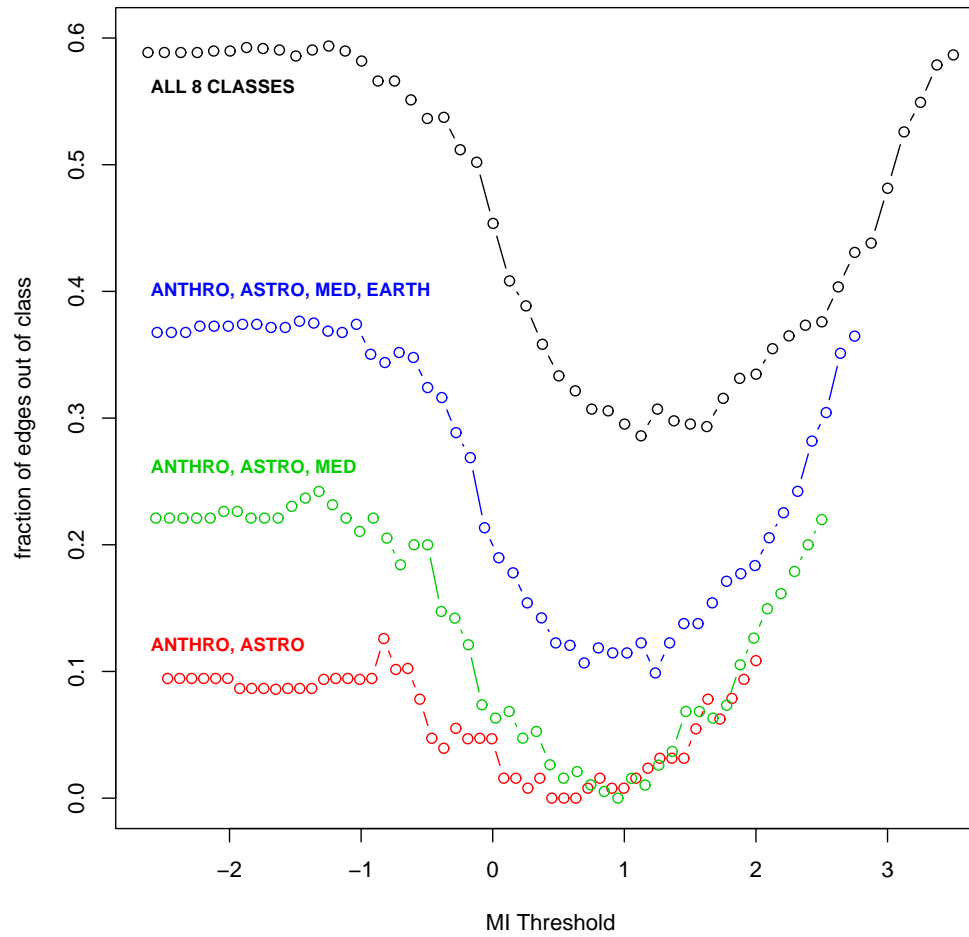
Graph Size = 300
Mutual Information Threshold = 0.5



27 edges between classes



Adding Documents to the Corpus



Future Work

- Optimal τ based on the "size" of the corpus
Unsupervised case
- Creating Random Documents
- Spectral Graph Analysis

