STAT992 Final Project:
Genre Identification with IMDB Data

Bi Cheng Wu, Chris Kardatzke

University of Wisconsin–Madison

December 10, 2020

Background

Summary from proposal

1. IMDB data of movies & cast (goal: identify genres)
2. Initial vsp clustering sensitive to country/language
3. Filtered out non-English titles; vsp still sensitive to cliques
4. New goal: mitigate influence of cliques to improve genre detection

Summary from proposal

1. IMDB data of movies & cast (goal: identify genres)
2. Initial vsp clustering sensitive to country/language
3. Filtered out non-English titles; vsp still sensitive to cliques
4. New goal: mitigate influence of cliques to improve genre detection

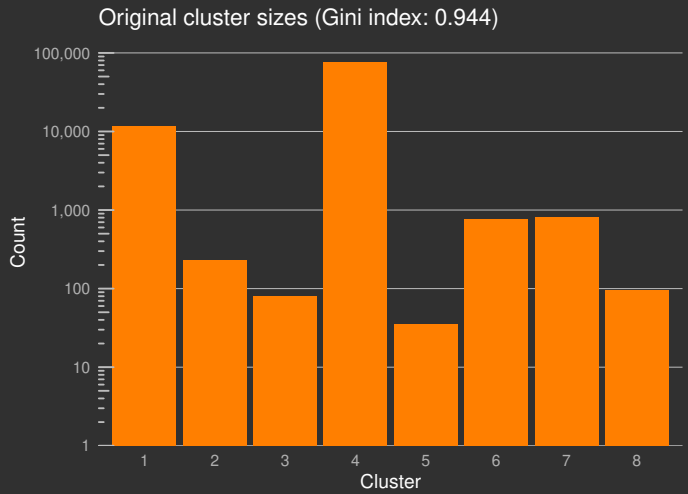
Summary from proposal

1. IMDB data of movies & cast (goal: identify genres)
2. Initial vsp clustering sensitive to country/language
3. Filtered out non-English titles; vsp still sensitive to cliques
4. New goal: mitigate influence of cliques to improve genre detection

Summary from proposal

1. IMDB data of movies & cast (goal: identify genres)
2. Initial vsp clustering sensitive to country/language
3. Filtered out non-English titles; vsp still sensitive to cliques
4. New goal: mitigate influence of cliques to improve genre detection

Previous clusters



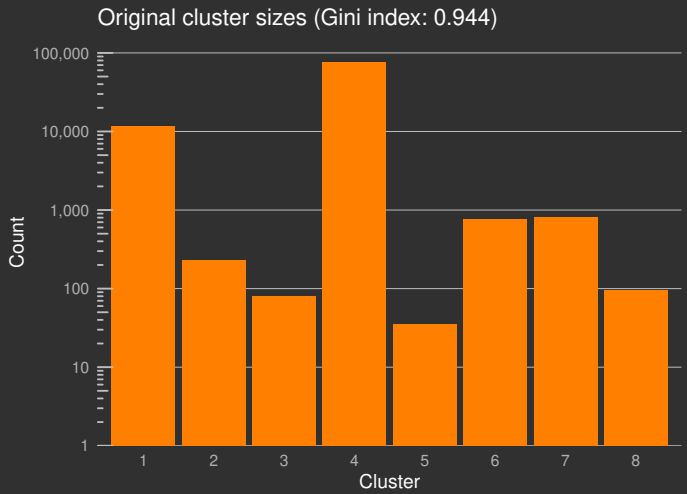
Some of the clusters:

2. H20 wrestling recordings
3. Blondie movies
4. Westerns
7. RiffTrax releases
8. Christian "documentaries"

$$\text{Gini} = \frac{1}{2n^2\bar{x}} \sum\sum |x_i - x_j|$$

(more equal) $0 \leq G \leq 1$ (less equal)

Previous clusters



Some of the clusters:

2. H20 wrestling recordings
3. Blondie movies
4. Westerns
7. RiffTrax releases
8. Christian "documentaries"

$$\text{Gini} = \frac{1}{2n^2\bar{x}} \sum\sum |x_i - x_j|$$

(more equal) $0 \leq G \leq 1$ (less equal)

Method 1: Predicting Node Cliquishness by Logistic Regression (LR)

Approach

1. Generate node statistics
2. Predict node “cliquishness” using logistic regression

Training Data

1. Clique nodes: All titles from cluster 3 and 5. Top 100 from 2 and 7.
2. Non-clique nodes: 670 titles from IMDB top/favorite lists.

Node Statistics

1. Degree
2. Coreness
3. No. triangles at each node
4. Degree distribution to other nodes
    1. Mean, mode
    2. Standard deviation
    3. Skew, kurtosis

Node Statistics

1. Degree
2. Coreness
3. No. triangles at each node
4. Degree distribution to other nodes
    1. Mean, mode
    2. Standard deviation
    3. Skew, kurtosis

Node Statistics

1. Degree
2. Coreness
3. No. triangles at each node
4. Degree distribution to other nodes
    1. Mean, mode
    2. Standard deviation
    3. Skew, kurtosis

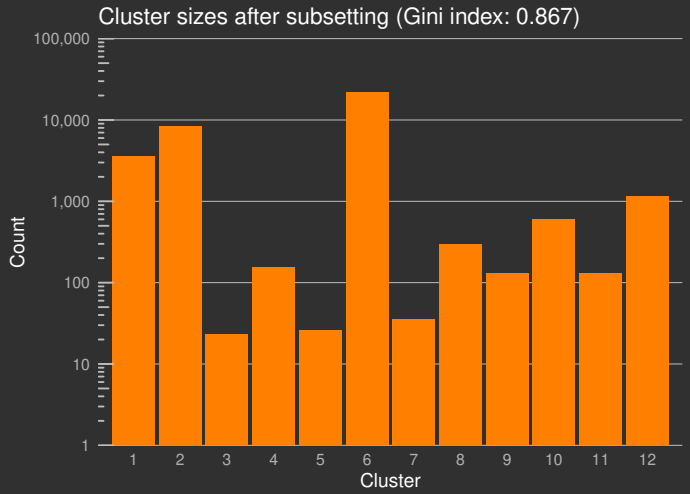
Node Statistics

1. Degree
2. Coreness
3. No. triangles at each node
4. Degree distribution to other nodes
    1. Mean, mode
    2. Standard deviation
    3. Skew, kurtosis

Node Statistics

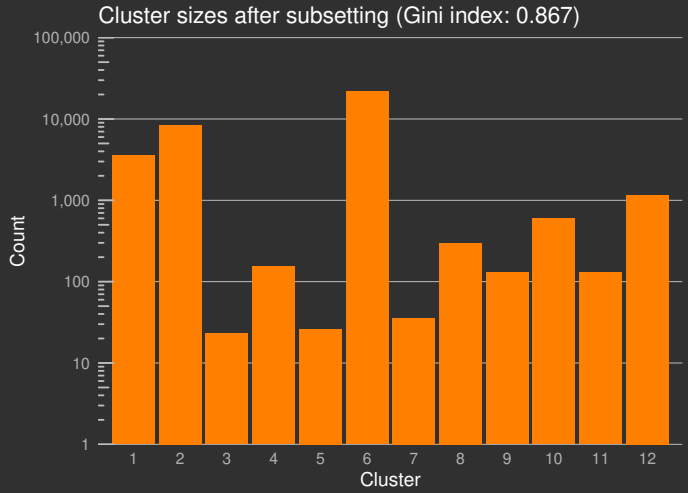
1. Degree
2. Coreness
3. No. triangles at each node
4. Degree distribution to other nodes
    1. Mean, mode
    2. Standard deviation
    3. Skew, kurtosis

Results



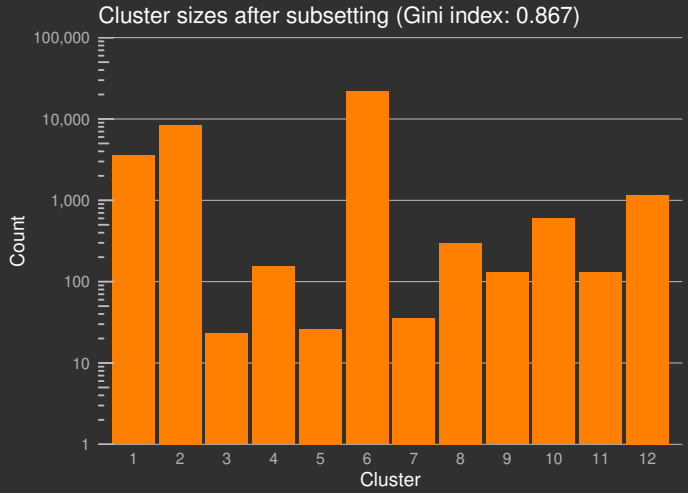
1. Gini slightly lower
2. $k = 8$ no longer best choice, use $k = 12$ instead
3. Clusters look much better
4. Link to results

Results



1. Gini slightly lower
2. $k = 8$ no longer best choice, use $k = 12$ instead
3. Clusters look much better
4. Link to results

Results

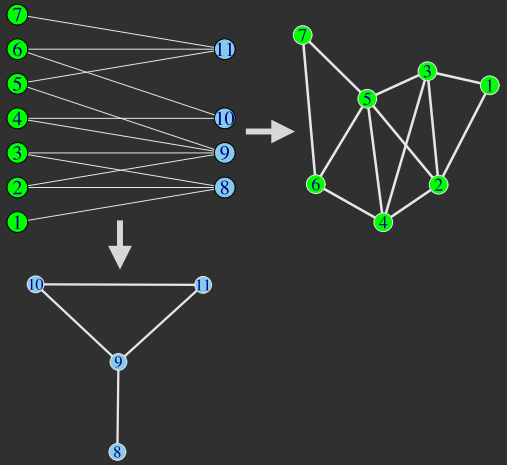


1. Gini slightly lower
2. $k = 8$ no longer best choice, use $k = 12$ instead
3. Clusters look much better
4. Link to results

Method 2: Transformation of Projected Adjacency

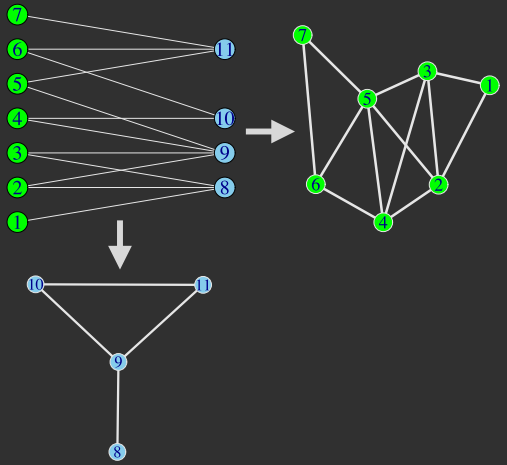
Bipartite graph projection

1. IMDB graph is bipartite
2. Project to titles $\times$ titles
3. Adjacency of projection may show cliques
4. Transformed adjacency may improve clusters
5. Used $\log_2(AA^T + 1)$



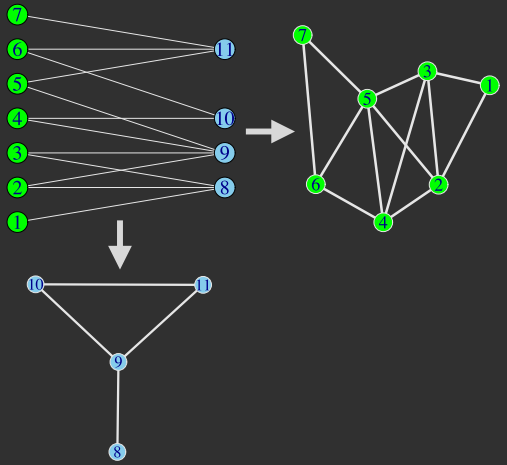
Bipartite graph projection

1. IMDB graph is bipartite
2. Project to titles $\times$ titles
3. Adjacency of projection may show cliques
4. Transformed adjacency may improve clusters
5. Used $\log_2(AA^T + 1)$



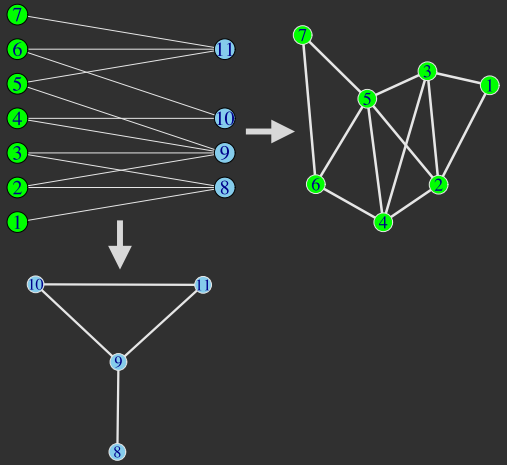
Bipartite graph projection

1. IMDB graph is bipartite
2. Project to titles $\times$ titles
3. Adjacency of projection may show cliques
4. Transformed adjacency may improve clusters
5. Used $\log_2(AA^T + 1)$

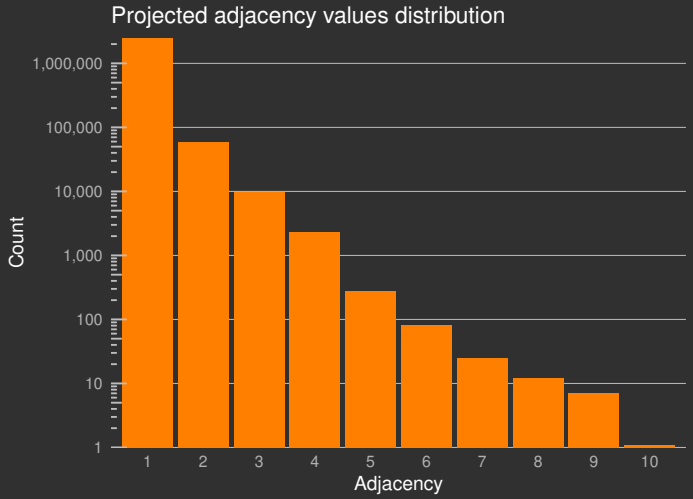


Bipartite graph projection

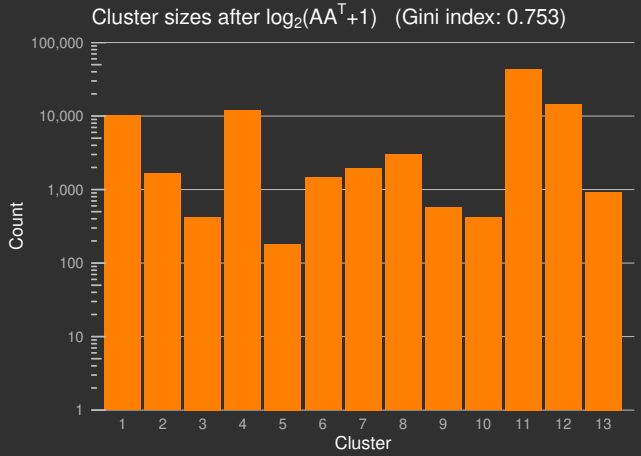
1. IMDB graph is bipartite
2. Project to titles $\times$ titles
3. Adjacency of projection may show cliques
4. Transformed adjacency may improve clusters
5. Used $\log_2(AA^T + 1)$



Projected IMDB adjacency matrix

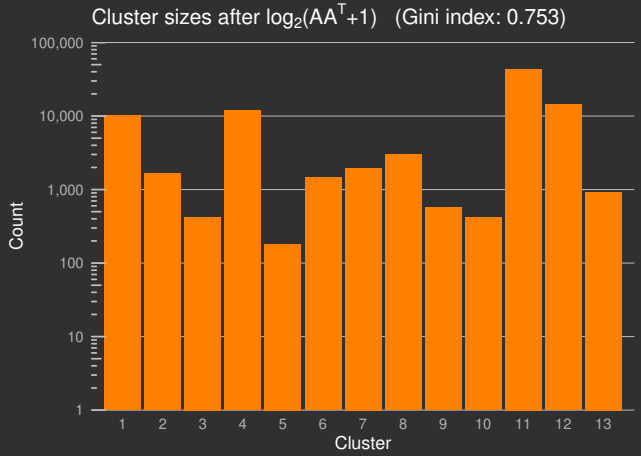


Results



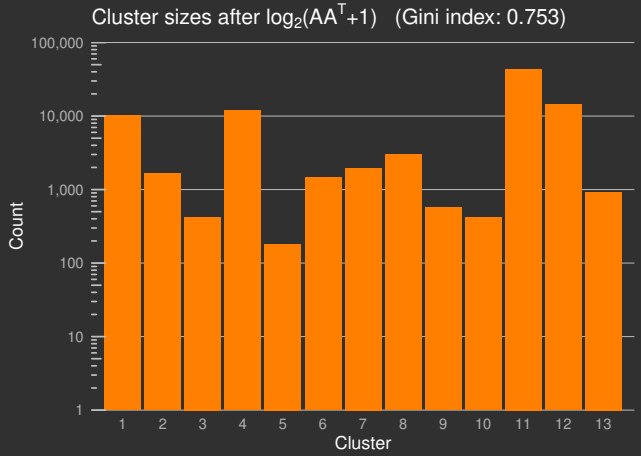
1. Gini much lower
2. chose $k = 13$
3. Link to results

Results



1. Gini much lower
2. chose $k = 13$
3. Link to results

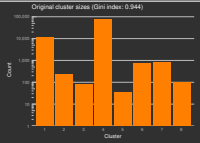
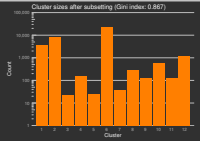
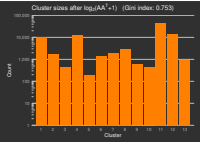
Results



1. Gini much lower
2. chose $k = 13$
3. Link to results

Summary

Results summary

Method	Clusters	Gini	Cluster quality
Original		0.944	bad, lots of cliques
Logistic regression		0.867	pretty good
Adjacency transform		0.753	also pretty good

Future work?

Ideas:

1. Use Personalized Page Rank (PPR) to identify cliquish nodes to contract
2. Scrape movies' descriptions, turn into bag of words, then run vsp

Ideas:

1. Use Personalized Page Rank (PPR) to identify cliquish nodes to contract
2. Scrape movies' descriptions, turn into bag of words, then run vsp

Acknowledgements

Special thanks to Karl Rohe for
guidance throughout the project