

Theatrical Genre Prediction using Social Network Metrics

Manisha Shukla, Susan Gauch,
and Lawrence Evalyn (2018)

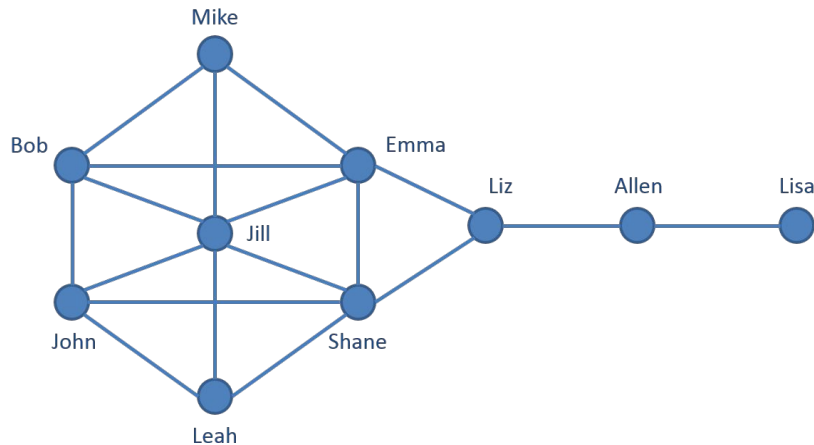
Presented by April Walker



Social Network Analysis (SNA)

Social network graph: a structure of **nodes** (representing an individual or character) and **edges** (representing relations amongst nodes)

SNA assumes that valuable information can be derived from the relationships between nodes





Social Network Analysis (SNA)

Previous unique utilizations of SNA:

- Detecting and disintegrating criminal social networks
 - Terrorist Activity ([Anggraini et al., 2015](#))
 - Money Laundering ([Dreżewski, et al., 2015](#))
- Studying the impact interpersonal football team relationships on game performance ([Trequattrini, et al., 2015](#))

This paper focuses on **genre prediction** using SNA. Because this method does not rely on the specific language of the play, it can be scaled extensively

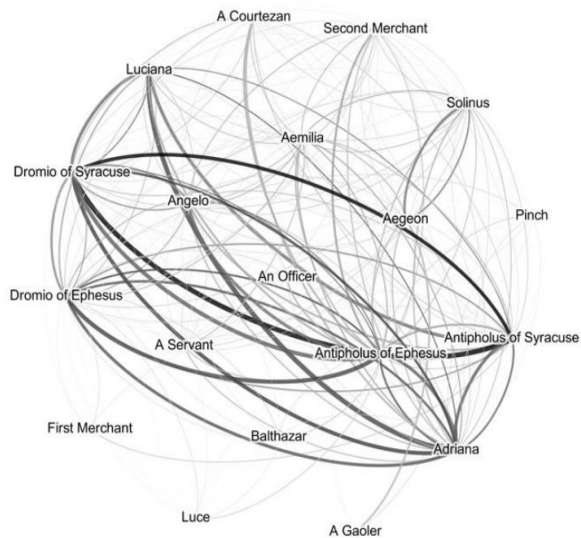


Motivation:

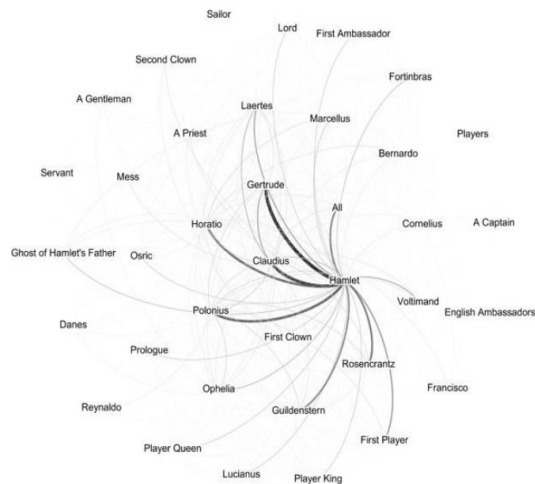
“...explore correlations between the mathematical properties of the social networks of plays and the plays’ dramatic genre...”

- Utilize social network theory as a method of feature extraction
- Study the importance of these features by testing their predictive power

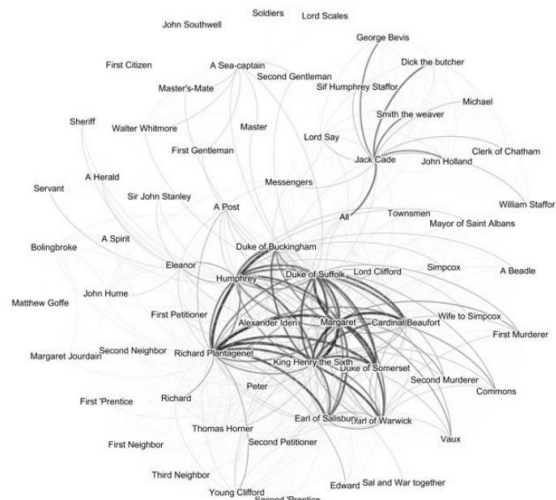
Social Network Graphs of Shakespearean Plays



Comedy of Errors



Hamlet



Henry V

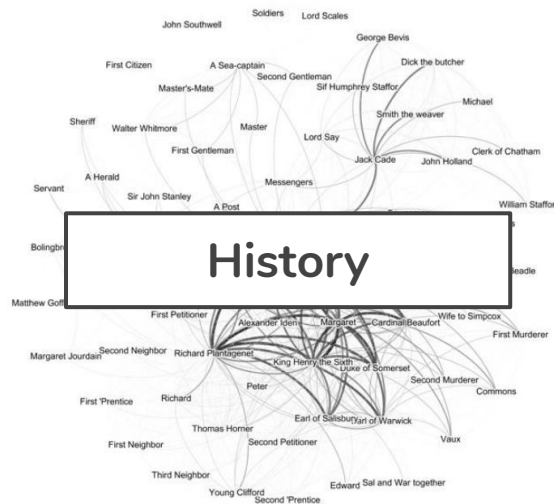
Social Network Graphs of Shakespearean Plays



Comedy of Errors

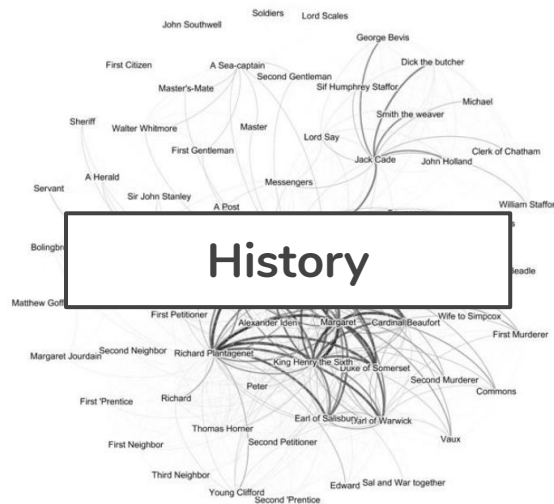
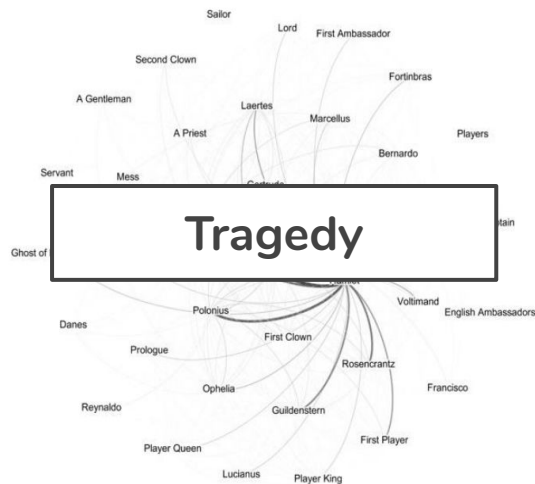


Hamlet



Henry V

Social Network Graphs of Shakespearean Plays



Additionally, this study hoped to analyze Shakespeare's disputed plays



Preprocessing:

TEI-encoded XML formatted Shakespearean plays were downloaded from [eXistdb Showcases](#)

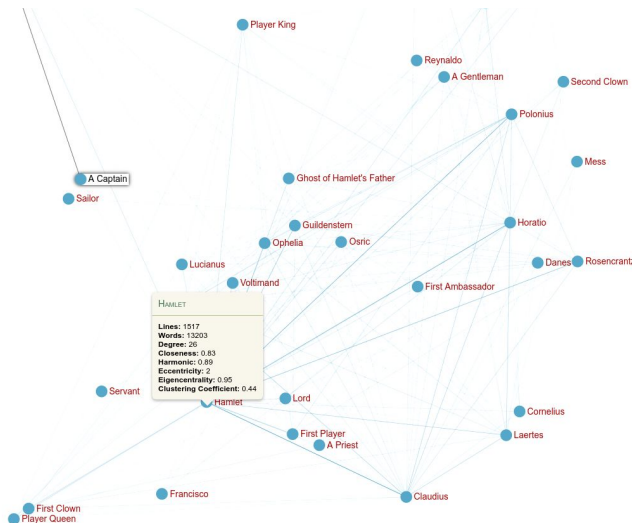
```
</sp>
<sp who="#F-ham-pol">
  <speaker rend="italic">Pol.</speaker>
  <l n="1206">Though this be madnesse,</l>
  <l n="1207">Yet there is Method in't: will you walke</l>
  <l n="1208">Out of the ayre my Lord?</l>
</sp>
<sp who="#F-ham-ham">
  <speaker rend="italic">Ham.</speaker>
  <l n="1209">Into my Graue?</l>
</sp>
<sp who="#F-ham-pol">
  <speaker rend="italic">Pol.</speaker>
  <l n="1210">Indeed that is out o'th'Ayre:</l>
  <l n="1211">How pregnant (sometimes) his Replies are?</l>
  <l n="1212">A happinesse,</l>
  <l n="1213">That often Madnesse hits on,</l>
  <l n="1214">Which Reason and Sanitie could not</l>
```

Extracted data to “Character” Java object:

- Scenes acted
- Number of lines spoken
- Number of words spoken
- Scene information to map other information



Social network graphs in addition to many social network metrics were calculated using [Gephi](#). All developed graphs are available on the UARK CSCE website [here](#).





Overview of Social Network Metrics

Node/Character Features:

Degree: number of adjacently connected nodes

Centrality: measures character importance based nodes relationship to other nodes (Eigenvector, Closeness, Harmonic)

Eccentricity: max number of connections to another node

Graph Features:

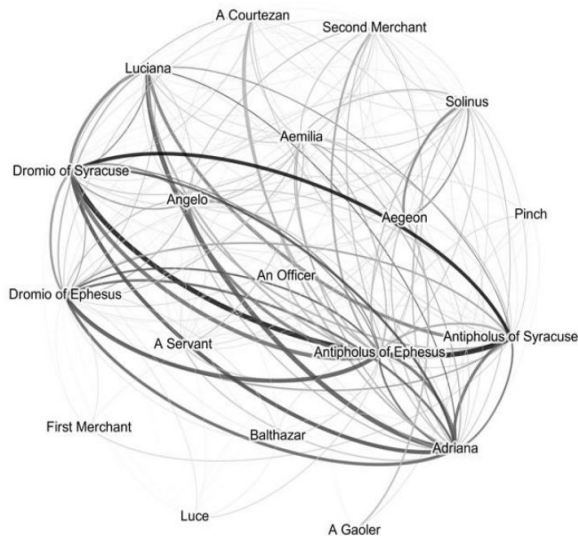
Number of characters (nodes), connections (edges), words, lines

Path Length: average distance between nodes

Graph Density: contrasts total connections with total possible

Diameter: max distance between nodes

Overview of Social Network Metrics



Graph Features:

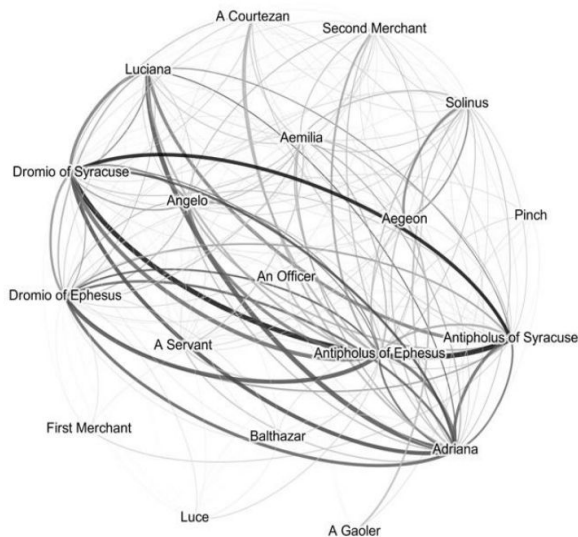
Number of characters (nodes), connections (edges), words, lines

Path Length: average distance between nodes

Graph Density: contrasts total connections with total possible

Diameter: max distance between nodes

Overview of Social Network Metrics



Each play was represented using **21 features**, 4 features being extracted from the text, **8 node/character features**, and **9 graph Features**

Example of Social Network Metric Output

	Total_No _Of_Ch aracters	Total_No _of_Edge s	Total_No _Of_Wor ds	Total_No _Of_Line s		Eigenvec tor	Eccentrici ty	Closenes s		Between ness	Clusterin g_Coeffi cient	Graph_D ensity	Diameter	Path_Len gth	Connecte d_Compo nents	Degree	Modulari ty	Weighte d_Degre e	Average_ Weighte d_Degre e	Radius	Class	
The Comedy of Errors	19	111	16431	1780	0.021944	0.243275	8.666667	6.902947	0.161008	0.007243	0.886123	0.649123	3	1.380117	1	0.330065	0.081421	765.8431	11.68421	9961.895	2	Comedy
The Merchant of Venice	23	107	23849	2689	0.046042	0.385156	8.666667	9.60768	0.220386	0.019932	0.813376	0.422925	3	1.656126	1	0.4329	0.123503	1430.714	9.304348	10305.3	2	Comedy
The Merry Wives of Windsor	24	186	23953	2726	0.014431	0.242326	0	8.426811	0.147448	0.004449	0.85839	0.673913	2	1.326087	1	0.3083	0.098	1019.984	15.5	12879.33	2	Comedy
The Taming of the Shrew	26	155	20883	2358	0.020523	0.414109	12	13.10841	0.234133	0.009851	0.860739	0.476923	3	1.544615	1	0.48	0.110488	1514.407	11.92308	10595.15	2	Comedy
The Tempest	19	114	18725	2273	0.017221	0.223931	9.333333	5.272727	0.123457	0.004277	0.858078	0.666667	3	1.368421	1	0.248366	0.220402	2015.621	12	11675.89	2	Comedy
The Winter's Tale	34	166	28182	3370	0.02238	0.520183	18.5	12.9393	0.211806	0.007375	0.791944	0.2959	4	1.882576	2	0.329545	0.322568	1003.203	9.764706	10001.71	0	Comedy
Twelfth Night or What You Will	18	101	22092	2548	0.063196	0.235338	1	7.812542	0.179931	0.024163	0.891659	0.660131	2	1.339869	1	0.382353	0.082598	1291.846	11.22222	13754.78	1	Comedy
Two Gentlemen of Verona	17	66	19312	2252	0.044561	0.360075	8	6.999725	0.213542	0.019681	0.830213	0.485294	3	1.551471	1	0.441667	0.093418	1077.183	7.764706	7362.706	2	Comedy
The First Part of King Henry VI	32	134	27286	3094	0.035597	0.569057	14	28.25026	0.383945	0.013014	0.83539	0.270161	4	2.013761	2	0.365591	0.35804	1271.155	8.375	7943.063	1	History
The First Part of King Henry VI	53	324	23945	2697	0.023478	0.592702	6.666667	25.37881	0.267505	0.011154	0.801583	0.235123	3	1.96807	1	0.475113	0.343134	396.5354	12.22642	4994.264	2	History
The Life and Death of King Henry VI	27	154	23301	2642	0.023727	0.435968	10.66667	14.79265	0.258383	0.012411	0.849965	0.438746	3	1.60114	1	0.523077	0.144977	1396.715	11.40741	11710.37	2	History
The Life of King Henry VI	45	202	26949	3168	0.018003	0.650109	24.4	19.94371	0.266839	0.008145	0.826716	0.20404	5	2.142424	1	0.452431	0.296406	730.8214	8.977778	6239.022	3	History
The Life of King Henry VI	44	190	29065	3296	0.03087	0.652198	26.8	21.1969	0.292599	0.014563	0.827164	0.200846	5	2.165006	2	0.471761	0.280177	1440.227	8.636364	6391.227	0	History
The Second Part of King Henry VI	49	213	29223	3314	0.019224	0.681401	24.57143	18.99207	0.295892	0.008479	0.820966	0.181122	7	2.582794	3	0.441046	0.392336	810.7207	8.693878	6043.755	0	History
The Second Part of King Henry VI	64	409	28062	3110	0.015384	0.643565	20	63.64174	0.441929	0.005853	0.854131	0.202877	4	2.159091	2	0.38044	0.241566	597.5428	12.78125	6913.219	1	History
The Third Part of King Henry VI	40	226	26890	2915	0.015577	0.531914	18.5	19.2841	0.264081	0.007442	0.831863	0.289744	4	1.897436	1	0.477733	0.201878	739.9825	11.3	7294.65	2	History
The Tragedy of King Richard III	31	148	24972	2794	0.03214	0.523905	13.5	16.155	0.274444	0.01655	0.754197	0.31828	4	1.834409	1	0.514943	0.135177	1311.707	9.548387	7958.581	2	History
The Tragedy of King Richard III	55	330	32689	3672	0.011765	0.632978	32	46.62182	0.387037	0.006783	0.838712	0.222222	4	1.853678	3	0.57652	0.168467	1524.881	12	8002.764	0	History
Antony and Cleopatra	53	296	27844	3510	0.014482	0.615835	19	47.10116	0.403526	0.007187	0.773574	0.214804	4	1.938088	2	0.496229	0.233219	743.6308	11.16981	5629.396	1	Tragedy
Coriolanus	52	295	30858	3750	0.025215	0.650693	8	43.1083	0.382019	0.017819	0.831559	0.222474	3	1.817292	2	0.706667	0.16565	1194.149	11.34615	7836.769	1	Tragedy
Cymbeline	39	191	31178	3727	0.04049	0.556134	14.66667	30.4393	0.349123	0.019323	0.874195	0.25776	3	1.804107	2	0.47724	0.262642	697.2745	9.794872	8182.359	1	Tragedy
Hamlet Prince of Denmark	34	192	34071	4051	0.022548	0.511784	12.66667	20.361	0.298439	0.01358	0.858524	0.342246	3	1.695187	1	0.602273	0.084707	2886.739	11.29412	11904.29	2	Tragedy
Julius Caesar	47	279	22034	2590	0.015789	0.628239	22.66667	34.25451	0.328847	0.010084	0.86208	0.258094	3	1.710671	2	0.638647	0.172884	1322.745	11.87234	6820.936	1	Tragedy
King Lear	25	154	29293	3481	0.014018	0.350111	11.33333	9.911793	0.192998	0.005511	0.812909	0.513333	3	1.513333	1	0.393116	0.048439	1827.92	12.32	12673.52	2	Tragedy
Macbeth	43	185	19052	2330	0.025166	0.656219	5.333333	23.77059	0.311413	0.014702	0.799068	0.204873	3	1.949059	1	0.584204	0.237471	754.2927	8.604651	3513.209	2	Tragedy



Genre Predictor

The popular Support Vector Machine was utilized to develop the classifier. In order to accommodate more than two classes, it was combined with One vs One (OvO) classification.

OvO Classification:

- Develop binary classification for each pair
- Chose class with most “votes”



Results

Feature selection was of high importance, as an SVM trained on all features only gave an accuracy of 66.43%

The SVM was instead trained on individuals, pairs, and triads of features

The top triad of features lifted this accuracy to 83.57%

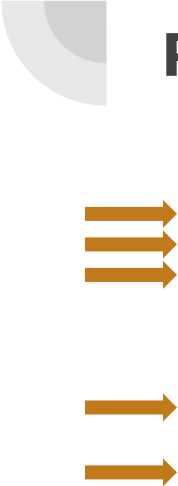
Results for single feature

SVM	
Feature	Accuracy
Path Length	66.43
Graph Density	61.07
Diameter	58.57
Characters	55.71
Eigenvector	55.71
Eccentricity	55.71
Harmonic	55.71
Average Weighted Degree	55.71
Lines	55.36
Degree	55.36
Closeness	52.50
Connected Components	50.35
Modularity	50.00
Words	47.50
Edges	47.14
Radius	47.14
Weighted Degree	44.28
Criticality	41.43
Clustering Coefficient	38.93
Average Degree	33.21
Betweenness	27.85

Table 2: Average feature value for each genre.

Features	Comedy	History	Tragedy
Characters	23.14	44	38.333
Edges	132	233	217.75
Words	22426.42	27238.2	27050.58
Lines	2586.5	3070.2	3215
Criticality	0.03	0.022	0.020
Eigenvector	0.34	0.59	0.52
Eccentricity	8.63	19.11	13.01
Closeness	9.28	27.42	24.95
Harmonic	0.19	0.31	0.29
Betweenness	0.01	0.010	0.011
Clustering Coefficient	0.84	0.82	0.83
Graph Density	0.52	0.25	0.34
Diameter	2.85	4.3	3.08
Path Length	1.516	2.02	1.71
Connected Components	1.07	1.7	1.5
Degree	0.37	0.46	0.52
Modularity	0.14	0.25	0.16
Weighted Degree	1306.85	1022.02	1457.85
Average Degree	11.31	10.39	11.38
Average Weighted Degree	11353.31	7349.09	9136.53
Radius	1.78	1.3	1.33

Results for single feature vs feature pairs




SVM	
Feature	Accuracy
Path Length	66.43
Graph Density	61.07
Diameter	58.57
Characters	55.71
Eigenvector	55.71
Eccentricity	55.71
Harmonic	55.71
Average Weighted Degree	55.71
Lines	55.36
Degree	55.36
Closeness	52.50
Connected Components	50.35
Modularity	50.00
Words	47.50
Edges	47.14
Radius	47.14
Weighted Degree	44.28
Criticality	41.43
Clustering Coefficient	38.93
Average Degree	33.21
Betweenness	27.85

SVM		
Feature 1	Feature 2	Accuracy
Harmonic	Diameter	72.50
Harmonic	Path Length	72.50
Graph Density	Diameter	72.50
Graph Density	Path Length	72.50
Lines	Path Length	72.14

Pictured above are the feature pairs which provided at least 70% accuracy in genre prediction using a SVM

Recall Eigenvector, Closeness, and Harmonic are all methods of measuring centrality

Results for single feature vs feature triads




SVM	
Feature	Accuracy
Path Length	66.43
Graph Density	61.07
Diameter	58.57
Characters	55.71
Eigenvector	55.71
Eccentricity	55.71
Harmonic	55.71
Average Weighted Degree	55.71
Lines	55.36
Degree	55.36
Closeness	52.50
Connected Components	50.35
Modularity	50.00
Words	47.50
Edges	47.14
Radius	47.14
Weighted Degree	44.28
Criticality	41.43
Clustering Coefficient	38.93
Average Degree	33.21
Betweenness	27.85

SVM			
Feature 1	Feature 2	Feature 3	Accuracy
Words	Characters	Lines	83.57
Words	Lines	Eigenvector	83.21
Words	Lines	Closeness	81.07
Lines	Eigenvector	Path Length	80.71
Lines	Harmonic	Path Length	80.71

Pictured above are the feature triads which provided at least 80% accuracy in genre prediction using a SVM

Recall Eigenvector, Closeness, and Harmonic are all methods of measuring centrality

Results for single feature vs feature triads



SVM	
Feature	Accuracy
Path Length	66.43
Graph Density	61.07
Diameter	58.57
Characters	55.71
Eigenvector	55.71
Eccentricity	55.71
Harmonic	55.71
Average Weighted Degree	55.71
Lines	55.36
Degree	55.36
Closeness	52.50
Connected Components	50.35
Modularity	50.00
Words	47.50
Edges	47.14
Radius	47.14
Weighted Degree	44.28
Criticality	41.43
Clustering Coefficient	38.93
Average Degree	33.21
Betweenness	27.85

SVM			
Feature 1	Feature 2	Feature 3	Accuracy
Words	Characters	Lines	83.57
Words	Lines	Eigenvector	83.21
Words	Lines	Closeness	81.07
Lines	Eigenvector	Path Length	80.71
Lines	Harmonic	Path Length	80.71

Additional numbers of feature combinations were not tested as computation time be wildin'



Results for problem plays

Table 16: Original and predicted classes for Problem plays

Play Name	Original Genre	SVM	Naïve Bayes
All's Well That Ends Well	Comedy	Comedy	Comedy
Measure for Measure	Comedy	Comedy	Comedy
Troilus and Cressida	Tragedy	Comedy	Tragedy



Conclusions

Shukla et al's methodology successfully classified plays without relying on the vocabulary itself, implying scalability across languages.

Future work:

- Reduce “false positives”
 - Make edges directional (capture relationship imbalance)
 - Incorporate NLP to distinguish targets of speech

Personal Questions:

- Explore additional feature selection methods and classifiers
- How scalable is this on a minimally processed corpus?

Questions?

Slides available at aprilwalker.io/NLP/presentation.pdf

Theatrical Genre Prediction using Social Network Metrics

Manisha Shukla, Susan Gauch,
and Lawrence Evalyn (2018)

Presented by April Walker