|  |
| --- |
| Theatrical Genre Prediction Using Social Network Metrics |
| Manisha Shukla1, Susan Gauch1 and Lawrence Evalyn2 |
| 1Department of Computer Science and Engineering, University of Arkansas, Fayetteville, AR, USA  *2Department of English, University of Toronto, Toronto, ON, Canada*  mshukla@email.uark.edu, sgauch@uark.edu, lawrence.evalyn@mail.utoronto.ca |

Keywords: Social Networks, Genre Prediction, Relationship Mining, Social Network Analysis, Network Theory.

Abstract: Networks study connections within large groups of objects. The anlaysis of how vertices are linked by edges has revealed many unexpected features of large systems. In this paper we apply network theory in the field of literature and explore how well social network features perform in literary analysis. We were able to successfully predict genre of Shakespeare’s plays with the help of social network metrics. Future work can be applied to fast and detailed literary analysis using social network metrics and network theory concepts.

# 1 INTRODUCTION

In literary studies, the three key areas of research could be defined as philology (the study of words), bibliography (the study of books as objects), and criticism (the evaluation or interpretation of literary meaning). Particularly since the advent of New Criticism, “the basic task of literary scholarship has been close reading of texts” [Moretti], which builds textual interpretations from precise study of specific words. Computational approaches to literature offer an alternate methodology for the work of literary study without close reading. “Distant reading” takes many forms, but for it to produce new literary insights, it must provide information about literary texts which is not easily accessible by reading them, and must do so for more texts than it is feasible for a person to read. Our paper presents a distant reading method which may aid in the task of literary criticism: using network graph analysis on social networks generated from the scripts of plays.

Social network analysis is well-established as a way to study social groups. In most approaches, the relationships between social actors are modeled as a graph, allowing a variety of new and existing graph algorithms to be applied. Social network analysis has been less often applied to literary works, but some scholars have created networks of plays wherein nodes represent characters, and edges represent interaction between pairs of characters in the play. Because these plays graphs are handmade for a very small number of plays, however, almost no work has been done to study the ability of mathematical properties of network graphs to predict features of a play. We address this gap by exploring correlations between the mathematical properties of the networks and dramatic genre.

Properties of social graphs are described with respect to two levels: ‘global graph metrics’ and ‘local graph metrics’. Global graph metrics consider the characteristic of the graph as a whole e.g., its diameter, mean node distance, betweenness, size of the giant component, clusters, small-worldness [8], etc., whereas the ‘local metrics’ relate to the features native to individual nodes such as degree, neighborhood, etc. Focusing on global graph metrics, we argue that the basic graph of a play reveals a great deal of information about the play and that social network metrics are, in fact, able to predict a play’s genre.

In this paper, we study the social networks of Shakepeare’s plays. To do so, we build a weighted, undirected graph of each play, in which characters are nodes, communication between characters represents an edge, and the weights represent the sum of total number of words spoken by two characters in the scenes in which they appear together. We define three classes for Shakespeare’s plays, namely ‘comedy’, history’ and ‘tragedy’. Then, we study the plays from the three groups to identify which features of their social networks allow us to distinguish between the three classes. With the class definitions and features in hand, we train a Support Vector Machine (SVM) classifier on our three classes.  We then use One Vs One (OvO) classification  to identify the top matching class for each play in our test set. We used TEI encoded XML formatted plays which we downloaded from [40]. Once the SVM is trained, it is used to predict the genre of unseen plays.

In a nutshell, our contributions are as follows: (1) we offer a list of individual and social factors that may help in literary analysis; and (2) we create a classifier to find the genre of an Early Modern play using the above individual and social metrics.

The rest of the paper is organized as follows. In Section 2, we present the existing work on social network analysis in different use cases. Section 3 describes our system. Section 4 contains experimental results, and Section 5 summarizes our findings and offers suggestions for possible future improvements.

# 2 Related Work

**2.1 Support Vector Machines**

Support Vector Machines (SVMs) are a popular machine learning method for classification, regression, and other learning tasks. Since our classification problem had more than two classes, we combined SVM with One vs One (OvO) classification. In this technique, we pick a pair of classes from a set of *n* classes and develop a binary classifier for each pair. So, given *n* classes, we can pick all possible combinations of pairs of classes from *n* and then for each pair we develop a binary SVM.

For combinations we have:

and for *k*=2 corresponding to two pairs we have

Therefore, we need C(3,2) = 3 SVM classifiers. These vote for each play’s class and the winning class gets picked as the play’s classification.

By using OvO, our SVM is much less sensitive to the problems of imbalanced datasets, which is particularly helpful given the different sizes of each of our three classes and our small overall sample size.

**2.2 Social Network Analysis**

A social network graph is a set of vertices and edges (called a sociogram) where vertices represent social actors and edges represent social relations among the vertices. However, a social network is more than just a set of vertices and lines, as its structure contains implicit information about the social actors and their relationships. The graph representation of social network offers a systematic and mathematical method for investigating these structures. Social network analysis is the process of investigating social network structures and ties through the use of network and graph theory concepts.

As [SOURCE] notes, “Social network analysis (SNA) is not a formal theory, but rather a wide strategy for investigating social structures,” which “borrows most of its core concepts from sociometry, group dynamics, and graph theory” [CITE].

In social network analysis of human activities, the nodes can be connected by many kinds of ties, such as “shared values, visions, and ideas; social contacts; kinship; conflict; financial exchanges; trade; joint membership in organizations; and group participation in events, among numerous other aspects of human relationships” [CITE]. Regardless of the nature of the connection, “[t]he defining feature of social network analysis is its focus on the structure of relationships” [CITE]. SNA methodologies assume that relationships between nodes are of central importance.

Social network analysis has been used in a wide variety of applications, with applications as diverse as disintegration models based on social network analysis of terrorist organizations [19], collaboration of scholars in graduate education [CITE], football team performance based on social network analysis of relaionships between football players [21], money laundering detection [22], and stress disorder symptoms and correlations in U.S. military veterans [23].

**2.3 Literary Analysis with SNA**

Because dramatic performances enact social encounters, social network analysis translates surprisingly well to fictional societies. Stiller et al. have shown that social networks in Shakespeare’s plays mirror those of real human interactions, particularly in size, clustering, and maximum degrees of separation (2003).

Surveying the field of literary analysis using SNA, Moretti categorizes several types of analyses: “an empirical, quantitative and hierarchical description of literary characters [31], corpus-based analyses exploring options for historical periodisation of literature [32] and types of aesthetic modelling of social formations in and by literary texts [33,34,35].” Moretti himself uses social networks to examine the plots of three Shakespearean tragedies, and to contrast a few chapters in English and Chinese novels (2011). Work following Moretti has focused on historical periodization, as in Algee-Hewitt’s examination of 3,439 plays looking only at the Gini Coefficient of each play’s eigenvector centrality to track changes in ensemble casts from 1500 to 1920 (2017). Our project focuses on a novel application, the classification of literary genre. When scaled up to a corpus covering a wider historical time span, our approach to genre could also provide insight on the historic periodization of literature. Moretti also identifies that, in the application of SNA to literature, “[m]ethods for the automated extraction of network data … and their evaluation are of particular importance [25, 26, 27, 28, 30, 31],” which we accomplish. [Moretti].

**2.4 Gephi Toolkit**

Gephi is an open source software for graph and network analysis, which allows for fast visualization and manipulation of large networks. As a generalist tool, “[i]t provides easy and broad access to network data and allows for spatializing, filtering, navigating, manipulating and clustering.” [Gephi Paper] Gephi also calculates a wide range of mathematical features for each graph, which we use as the basis for our mathematical analysis (as discussed in more detail in 3.3).

**3 Our Design**

Our system for identifying genre consists of three building blocks: the Play Parser, the Social Network Generator and the Genre Predictor. Figure 1 diagrams the main components of the system architecture, which are discussed in more detail in the following subsections.

Play Parser

Social Network Generator

Genre Predictor

Figure 1: Block diagram of our system.

**3.1 Play Parser**

The main purpose of this module is to parse each play for the basic information which will be the buildging blocks for the graph of the play. The information extracted includes the total number of characters in the play, the name and role of each character, the total number of acts and scenes, who was present in each scene (using stage directions to account for entrances and exits within scenes), who spoke, and how many lines and words are spoken. Table 2 shows some details.

Table 1: TEI encoded XML file information.

|  |  |  |
| --- | --- | --- |
| **XML Tag** | **Information contained** | **Example** |
| <casteItem> | Character name and role in the play | <castItem type="role">  <role xml:id="Pedro"> Don Pedro</role> <roleDesc> prince of Arragon </roleDesc>  </castItem> |
| <div> | Act | <div xml:id="sha-man1">  <head>Act 1</head> |
| <div> | Scene | <div xml:id="sha-man101">  <head>Act 1, Scene 1</head> |
| <speaker> | Current speaker | <speaker>Beatrice </speaker> |
| <l>, <ab> | Line | <l xml:id="sha-man101299" n="299">  And tell fair Hero I am Claudio,</l>  <ab xml:id="sha-man201311" n="311"> born to speak all mirth and no matter.</ab> |

**3.2 Social Network Generator**

This module builds play graphs using the information generated by the play parser in the previous section. In order to generate files which can be used to create graphs for plays, we used the Gephi API to generate gexf files. Each gexf file maps character to a node and communication between characters as an edge. Each character stores as an attribute the total number of lines and words spoken by that character in the play. After this mapping, each edge is weighted with the sum of total number of lines and words spoken by the two characters in their shared scenes. These form the basis of the extracted features in Table 2. Once the basic structure is ready, using the inbuilt functions of the Gephi API we calculate 17 metrics of the graph’s graphs features. These are the network features in Table 2.

**3.2.1 Extracted Features**

As extracted features, we chose to use the most simple and easily quantifiable metrics, such as total number of characters in the play (see Table 2). As our results in 4.3.1 and 4.3.3 demonstrate, despite their simplicity as features, the number of edges in the graph and the number of words spoken can play a crucial role in determining the genre.

**3.2.2 Network Features**

We compute the network features of each play solely from the generated graph. For features that describe an individual node, such as degree or EigenVector, we calculate the average value of that feature for all the nodes in the graph.

Table 2: Features extracted from Shakespeare’s plays. Here n represents a play in graph, c character node in a graph, and e an edge in graph.

|  |
| --- |
| **Extracted Features** |
| ***Features*** |
| **1. tot\_characters** = total number of characters of **n** |
| **2. tot\_edges** *=* total number of edges of **n** |
| **3. tot\_lines** = total number of lines spoken by **c** in **n** |
| **4. tot\_words** = total number of words spoken by **c** in **n** |
| **Network Features** |
| **5. Degree** = set of adjacent nodes of **c** in the graph |
| **6. Criticality** = A **k-critical graph** is a critical graph with chromatic number k; a graph G with chromatic number k is **k-vertex-critical** if each of its vertices is a critical element. |
| **7. EigenVector** = A measure of node importance in a network based on a node's connections. |
| **8. Eccentricity** = The **eccentricity of a graph** vertex in a connected **graph** is the maximum **graph** distance between and any other vertex of. |
| **9. Closeness** **Centrality** = The average distance from a given node to all other nodes in the network. |
| **10. Harmonic** **Centrality** = In a (not necessarily connected) graph, the **harmonic centrality** reverses the sum and reciprocal operations in the definition of closeness centrality. |
| **11. Betweeness** **Centrality** = Node Betweenness Centrality measures how often a node appears on shortest paths between nodes in the network. |
| **12. Clustering Coefficient** = The clustering coefficient (Watts-Strogatz), when applied to a single node, is a measure of how complete the neighborhood of a node is. When applied to an entire network, it is the average clustering coefficient over all nodes in the network. |
| **13. Density** = Measures how close the network is to complete. A complete graph has all possible edges and density equal to 1. |
| **14. Diameter** = The maximal distance between all pairs of nodes. |
| **15. Path Length** = The average graph-distance between all pairs of nodes. |
| **16. Connected Components** = A **connected component** of an undirected graph is a maximal set of nodes such that each pair of nodes is connected by a path. |
| **17. Modularity** = Measures how well a network decomposes into modular communities. |

## 3.3 Genre Predictor

The genre predictor is a support vector machine binary classifier which uses one vs one technique for classification. The model is trained using the extracted and network features of the 31 plays in the training dataset, listed in Table 4. It was then used to predict the genres of 6 plays in the testing dataset, two from each genre, listed in Table 5. This process generated a confusion matrix for each feature or combination of features tested. For example, the confusion matrix in table below shows the original and predicted genre of each play as determined by the three features of Lines, Eigenvector and Density.

Table 3: Genre Predictor results when trained using Lines, Eigenvector and Density as a set of Play Features.

|  |  |  |
| --- | --- | --- |
| **Play** | **Original Genre** | **Predicted Genre** |
| The Life and Death of King John | History | History |
| The Tragedy of King Richard the Second | History | History |
| Timon of Athen | Tragedy | Tragedy |
| Macbeth | Tragedy | Tragedy |
| The Comedy of Errors | Comedy | Comedy |
| Much Ado About Nothing | Comedy | Comedy |

# 4 Experiment

## 4.1 Dataset

Our dataset consisted of 37 plays by Shakespeare, in TEI encoded XML. XML format was chosen as it was much easier to fetch the information required from the play along with maintaining accuracy in the extraction. The dataset was downloaded from [40].

For our dataset, we used 31 plays as a training set, and remaining 6 as a test set. The test set conatined two plays from each genre. For 5 fold cross-validation we used all 37 plays. For a validation set, we downloaded six plays other than Shakespeare’s to validate how well our model performed in general.

Table 4: Training dataset.

|  |  |
| --- | --- |
| **Play** | **Genre** |
| All's Well That Ends Well | Comedy |
| As You Like It | Comedy |
| Cymbeline | Comedy |
| A Midsummer Night's Dream | Comedy |
| Love's Labour's Lost | Comedy |
| The Merchant of Venice | Comedy |
| Measure for Measure | Comedy |
| The Merry Wives of Windsor | Comedy |
| The Taming of the Shrew | Comedy |
| Pericles, Prince of Tyre | Comedy |
| Troilus and Cressida | Comedy |
| Two Gentlemen of Verona | Comedy |
| Twelfth Night or What You Will | Comedy |
| The Winter's Tale | Comedy |
| The First Part of King Henry the Fourth | History |
| The Life of King Henry the Fifth | History |
| The Life of King Henry the Eight | History |
| The First Part of King Henry the Sixth | History |
| The Second Part of King Henry the Fourth | History |
| The Second Part of King Henry the Sixth | History |
| The Third Part of King Henry the Sixth | History |
| The Tragedy of King Richard the Second | History |
| Julius Caesar | Tragedy |
| King Lear | Tragedy |
| Hamlet,Prince of Denmark | Tragedy |
| Macbeth | Tragedy |
| Othello | Tragedy |
| Romeo and Juliet | Tragedy |
| Titus Andronicus | Tragedy |
| Timon of Athens | Tragedy |
| Antony and Cleopatra | Tragedy |
| Coriolanus | Tragedy |

Table 5: Test dataset.

|  |  |
| --- | --- |
| **Play** | **Genre** |
| The Life and Death of King John | History |
| The Tragedy of King Richard the Second | History |
| Timon of Athen | Tragedy |
| Macbeth | Tragedy |
| The Comedy of Errors | Comedy |
| Much Ado About Nothing | Comedy |

## 4.2 Experimental Setup

We then use our generated network graphs to test our central question: whether the social network enacted by a play’s characters can be used as a proxy for features of the play’s narrative content. More specifically, we ask whether social networks can be used to distinguish between the dramatic genres of tragedy, comedy, and history. We tested 17 different mathematical features of the networks. We first tested how well individual features were able to distinguish between different genres. Our second test tried various pairs of features, and the third tried combinations of three features.

## 4.3 Results

The following table shows the average value for each network metric per genre.

Table 5: Average feature value for each genre.

|  |  |  |  |
| --- | --- | --- | --- |
| **Average Feature Value** | **Tragedy** | **History** | **Comedy** |
| Characters | 59.4 | 62.9 | 43.35 |
| Edges | 870.4 | 924 | 585 |
| Words | 31777.4 | 36544.1 | 33884.88 |
| Lines | 3774.7 | 4102 | 3936.94 |
| Criticality | 0.27 | 0.68 | 0.23 |
| Eigenvector | 0.39 | 0.39 | 0.52 |
| Eccentricity | 2.32 | 3.03 | 2.15 |
| Closeness | 0.56 | 0.52 | 0.62 |
| Harmonic | 0.62 | 0.57 | 0.68 |
| Betweenness | 12.53 | 19.24 | 7.45 |
| Clustering Coefficient | 0.73 | 0.74 | 0.81 |
| Graph Density | 0.48 | 0.45 | 0.61 |
| Diameter | 3 | 3 | 3.3 |
| Path Length | 1.51 | 1.60 | 1.39 |
| Connected Components | 2 | 1 | 1.23 |
| Degree | 28.13 | 28.45 | 26.27 |
| Modularity | 0.14 | 0.20 | 0.12 |

**4.3.1 Single Feature Accuracy**

Our first test attempted to identify genre using only a single feature at a time. However, no single feature was independently sufficient to identify the genre. Of the features tested, graph density provided the greatest accuracy (83%) for genre identification using a single feature.

Table 6: Genre prediction accuracy using a single feature.

|  |  |
| --- | --- |
| **Feature** | **Accuracy** |
| Graph Density | 83.33 |
| Total No. of Characters | 66.67 |
| Total No. of Lines | 66.67 |
| Path Length | 66.67 |
| Connected Components | 66.67 |
| Average Degree | 66.67 |
| Total No. of Edges | 50 |
| Total No. of Words | 50 |
| Average Criticality | 50 |
| Average Eccentricity | 50 |
| Average Closeness Centrality | 50 |
| Average Clustering Coefficient | 50 |
| Modularity | 50 |
| Average Eigenvector Centrality | 33.33 |
| Average Harmonic Centrality | 33.33 |
| Average Betweenness Centrality | 33.33 |
| Diameter | 33.33 |

**4.3.2 Pair of Features Accuracy**

However, if features are used in pairs, the network graphs can achieve full accuracy. Table 7 shows the pairs which were able to identify genre with 100% accuracy on our test set.

Table 7: Pairs of features which provided 100% accuracy in genre predection.

|  |  |
| --- | --- |
| **Feature 1** | **Feature 2** |
| Edges | Connected Components |
| Lines | Eccentricity |
| Eccentricity | Connected Components |
| Density | Connected Components |
| Modularity | Connected Components |

**4.3.3 Three Features Accuracy**

If we combine three features, the network graphs again achieve full accuracy. Table 8 shows the triads which were able to identify genre with 100% accuracy on our test set.

Table 8: Sets of three features which provided 100% accuracy in genre prediction.

|  |
| --- |
| **Feature 1, Feature 2, Feature 3** |
| Characters, Edges, Connected Components |
| Characters, Lines, Eigenvector |
| Characters, Eccentricity, Path Length |
| Characters , Density, Path Length |
| Characters, Connected Components, Degree |

**4.3.4 Discussion**

The relevance of graph density in distinguishing genres is visually obvious when individual comedy and history networks are compared.

(2henry6.png).

Figure 3: Network graph of *The Second Part of King Henry The Sixth*, a history.

(comedyerrors.png).

Figure 3: Network graph of *The Comedy of Errors*, a comedy.

(hamlet.png, caesar.png)

Figure 4: Network graphs of *Hamlet* and *Julius Caesar*, both tragedies.

Histories feature highly dispersed networks, with large numbers of very minor characters, such as “First,” “Second,” and “Third” members of groups like soldiers and ambassadors (Figure 2). Comedies, in contrast, feature networks with far fewer characters, in which nearly everybody speaks to nearly everybody else at some point (Figure 3). Graph density is insufficient, however, to fully distinguish the tragedies, which feature networks somewhere between history and comedy in their density, and show more variety overall (Figure 4). Therefore, more complex metrics are needed in combination with each other to accurately identify the three genres.

Table 7 shows that certain pairs of metrics are successful in capturing the diffrences completely for different genres. Table 8 shows that each metric captures a specific kind of information about the play and hence, the set of three pairs which provide 100% accuracy do not necessarily always include Table 5 pairs.

[Somewhere in the discussion we should mention the DH2018 paper and just have a sentence or two saying that this work has been used to examine the genre attributions of Shakespeare’s “problem plays”]

**5 Conclusion**

In this paper, we successfully classify plays based on their genre without using the actual words of the plays. Our networks of the well-studied works of Shakespeare can provide a baseline against which to contextualize similar studies of other plays. The network graphs themselves provide a new insight into the plays (the hidden shape of social relationships between characters). The application of mathematical graph analysis to these networks provides a dramatically faster way to determine important information about them (their genre).

**6 Future Work**

Since the parser is highly extensible and can be used with any plays encoded in TEI, future work applying these methods to literary analysis does not need to be restricted to plays that are similar to Shakespeare’s, but could be used to compare plays over a long period of time. Future work doesn’t even need to be be restricted to plays written in English; one future application in development, for example, will study eighteenth century plays written in English, French, and German.

Future refinements to the social network generator could make edges between nodes directional, to better capture imbalanced relationships between characters; this level of detail was not necessary to distinguish between Shakespeare’s plays, but might be important for different identification tasks. Natural Language Processing (NLP) could also be integrated into the parser to more accurately identify the targets of speech, to capture instances where characters are on stage but cannot hear what is being said or are not being spoken to. These kinds of improvements would reduce “false positives” in the creation of edges between nodes, perhaps enabling better analysis of larger or more complicated groups of literary plays.

**References**

References and citations should follow the Harvard (Author, date) System Convention (see the References section). As example you may consider the citation (Smith, 1998). Besides that, all references should be cited in the text. No numbers with or without brackets should be used to cite or to list the references.

References should be set to 9-point, justified, with a single linespace and hanging indent of 0,5-centimenter.

Citations should follow the section text format (see file FormatContentsForAuthors.pdf).

Moore, R., Lopes, J., 1999. Paper templates. In *TEMPLATE’06, 1st International Conference on Template Production*. SCITEPRESS.

Smith, J., 1998. *The book*, The publishing company. London, 2nd edition